Universitatea Politehnica Bucureşti
Facultatea de Automatică şi Calculatoare
Departamentul de Automatică şi Ingineria Sistemelor

Teză de doctorat

Analiza patogenezei moleculare a cancerului de sân folosind analiza integrată de microarray și modelare genică

Conducator de doctorat:
Prof.dr.ing. Cătălin Buiu

Doctorand:
Ing. Irina-Oana Lixandru-Petre

București, 2022

Cuprinsul rezumatului tezei de doctorat

Cuprinsul tezei de doctorat 1
Cuvinte cheie 4
Capitolul 1. Introducere 4
Capitolul 2. Expresia genelor 7
2.1. Bioinformatică și data mining 7
2.2. Reglarea expresiei genelor 7
2.3. Tehnici de profilare a expresiei genelor 10
Capitolul 3. Prezentarea sintetică a contribuțiilor științifice 12
3.1. Cadrul de implementare 12
3.2. Metodologie 13
Capitolul 4. Elemente de verificare și validare. Analiza critică a rezultatelor 25
Capitolul 5. Sinteza contribuțiilor științifice aduse și perspective de dezvoltare viitoare 28
5.1. Sinteza contribuțiilor științifice aduse 28
5.2. Perspective viitoare 36
Bibliografie selectivă 37

Cuprinsul tezei de doctorat

Acronime 1

Capitolul 1. Introducere 5

Capitolul 2. Contextul cercetării, domeniul de aplicabilitate și provocări ale cadrului global 8

2.1. Bioinformatică și Data Mining 8

2.2. Reglarea expresiei genelor 10

2.3. Alterări în reglarea expresiei genelor 13

2.4. Baze de date biologice 17

2.5. Formate biologice şi ontologii 19

2.6. Tehnici de profilare a expresiei genelor 20

2.7. Provocări în analiza expresiilor de gene 23

Capitolul 3. Analiza cadrului existent: diferite metode de analiză ale proceselor biologice şi expresiilor de gene 25

3.1. Modelare și modele 25

3.2. Tehnici inteligente 27

3.3. Analiza expresiei genelor 29

3.3.1. Alinieri ale expresiilor de gene 29

3.3.2. Analiza expresiilor de gene pe baza modelului probabilistic Bayes și a rețelelor bayesiene 31

3.3.3. Analiza expresiilor de gene cu ajutorul modelului Markov ascuns (HMM) 34

3.3.4. Analiza expresiilor de gene cu ajutorul metodelor celor mai mici pătrate (PLS) și a componentelor principale (PCA) 36

3.3.5. Analiza expresiilor de gene cu ajutorul grupării ierarhice 37

Capitolul 4. Problematica, caracteristicile și pasii generali ai unei analize de microarray 40

4.1. Achiziția datelor 43

4.2. Preprocesarea datelor și controlul calității 44

4.2.1. Data cleaning (reducerea zgomotului, inconsistența datelor, valori lipsă) 44

4.2.2. Explorarea datelor (Data exploration) 45

4.3. Reducerea datelor (Data reduction and transformation) prin eliminarea atributelor redundante si transformarea datelor într-un format comun 46

4.4. Determinarea genelor diferențiate 48

4.5. Alegerea modelului 51
4.6. Validare model. Evaluarea performanței 54
4.7. Vizualizare și adnotare gene 59
4.8. Integrarea datelor (Data integration) 59
Capitolul 5. Contribuții științifice. Obiective urmărite în cadrul tezei și specificații de realizare ale acestora. Metodologie și cadrul de implementare. 61
5.1. Cadrul de implementare 61
5.2. Abordarea propusă în analiza expresiilor de gene 63
5.2.1. Partea I. Descărcarea datelor biologice, controlul calității, normalizarea datelor, reducerea dimensionalității, gruparea ierarhică și selectarea genelor exprimate diferențial. 64
5.2.1.1. Pas 1. Descărcarea datelor biologice 64
5.2.1.2. Pas 2. Controlul calității si normalizarea datelor 66
5.2.1.3. Pas 3. Reducerea dimensionalității 68
5.2.1.4. Pas 4. Gruparea ierarhică și selectarea genelor exprimate diferențial 69
5.2.2. Partea a II-a. Gruparea genelor corelate în module de co-expresie, selectarea genelor cu cele mai mari conectivități intramodulare (gene hub), adnotarea genelor, realizarea de rețele de gene pentru hub-uri. 71
5.2.2.1. Pas 1. Gruparea genelor corelate în module de co-expresie 71
5.2.2.2. Pas 2. Selectarea genelor cu cele mai mari conectivități intramodulare (gene hub) 74
5.2.2.3. Pas 3. Adnotare gene 74
5.2.2.4. Pas 4. Realizarea de grafuri de co-expresie sau rețele de gene pentru hub-uri 75
5.2.3. Partea a III-a. Identificarea genelor corelate cu diferite tipuri de cancer, analiza genelor corelate cu cancerul de sân, analiza genelor din modulul greenyellow. 79
5.2.3.1. Pas 1. Identificarea genelor corelate cu diferite tipuri de cancer 79
5.2.3.2. Pas 2. Analiza genelor corelate cu cancerul de sân 83
5.2.3.3. Pas 3. Analiza genelor din modulul greenyellow 86
5.2.4. Partea a IV-a. Modelarea genelor din modulul greenyellow cu ajutorul rețelelor bayesiene, diferite interpretari ale rezultatelor inferențelor probabilistice 95
5.2.4.1. Pas 1. Modelarea genelor din modulul greenyellow 95
5.2.4.2. Pas 2. Interpretarea separată a rezultatelor inferenței probabilistice asupra tipurilor de conexiuni din rețeaua bayesiană formată din genele CLEC12A, STK17A și ADAM28 (cazul general) 101
5.2.4.3. Pas 3. Interpretarea rezultatelor inferenței probabilistice asupra rețelei bayesiene alese 105
5.2.4.4. Pas 4. Studiu de caz. Lipsa influenței genei CLEC12A asupra genei STK17A și interpretarea rezultatelor inferenței probabilistice asupra rețelei bayesiene nou rezultate 108

5.3. Observații 111

Capitolul 6. Elemente de verificare și validare: scenarii de experimentare și rezultate obținute. Analiza critică a rezultatelor și demonstrarea atingerii obiectivelor 113

Capitolul 7. Sinteza contribuțiilor științifice aduse, concluzii, elemente de originalitate și perspective de dezvoltare viitoare 128

7.1. Sinteza contribuțiilor științifice aduse 128

7.2. Perspective viitoare 137

ANEXA 139

Bibliografie 140

Bilanțul producției științifice și a activității de cercetare 152

Cuvinte cheie

Gene ; microarray ; reţele bayesiane ; cancer ; mutaţii ;

Capitolul 1. Introducere

Bioinformatica este ştiinţa aflată la intersecţia unor domenii variate precum biologia, medicina, matematica sau știința calculatoarelor. Definită ca disciplina care se referă la analiza şi interpretarea datelor biologice cu scopul obţinerii de informaţii şi cunostiinţe, această ramură își propune să ajungă la o înţelegere cât mai profundă a mecanismelor vieţii, la oricare nivele.

Sistemele biologice genice sunt caracterizate de relații continue de comunicare cu alte celule sau gene, alterările din acest sistem complex de comunicaţie putând avea urmări grave asupra stabilităţii celulare, crescând riscul promovării unor boli periculoase asupra sănătăţii organismelor. Printre aceste boli se numără şi cancerul, o modificare a echilibrului celular, caracterizată printr-un număr ridicat de celule alterate, care se divid infinit şi formează tumori care pot să invadeze întreg organismul.

Lucrarea de faţă abordează analiza de microarray, printr-o tehnică integrată de interpretare a unor expresii genice provenite din țesut mamar modificat, în scopul unei mai bune înțelegeri a modului prin care mutațiile se pot transmite de la o genă la alta, ducând la formarea celulelor canceroase care deteriorează întreg organismul.

Cancerul este şi va rămâne o problemă mondială privind alterarea sănătăţii indivizilor. Deşi s-a ajuns la un grad de cunoaştere ridicat al informațiilor, specialiştii nu au reuşit încă să găsească răspuns la toate procesele și modurile de comunicare care au loc, înainte ca o celulă să sufere mutaţii într-atât de multe şi diferite, încât să dea naştere unor clone ce promovează formarea ţesuturilor neoplazice. Motivaţia alegerii acestei teme este deci implicaţia majoră pe care cancerul o are în cadrul sistemului public de sănătate şi modul în care acesta acţionează de la un nivel micro, de mutaţie a unei celule, până la un nivel macro, de formare a metastazelor,

reuşind să treacă cu succes de barierele pe care corpul le ridică, în vederea menţinerii stabilităţii celulare.

Teza este structurată în șapte capitole, urmate de o anexă ce cuprinde codul din limbajul de programare R și partea de bibliografie, ce a stat la baza informațiilor teoretice din cadrul acestei lucrări.

Capitolul 2 include contextul cercetării și domeniul de aplicabilitate, reprezentând o introducere în domeniile bioinformatică și data mining. Acesta începe cu prezentarea conceptelor biologice de bază, legate de materialul genetic și reglarea expresiilor de gene, continuând cu expunerea principalelor baze de date și formate biologice existente. În continuare, sunt prezentate pe larg principiile a două tehnologii de microarray des folosite și anume cele pe bază de acid dezoxiribonucleic complementar și oligonucleotide, capitolul încheindu-se cu prezentarea mai multor provocări computaționale, legate de analiza expresiilor de gene.

Capitolul 3 include prezentarea cadrului existent legat de analiza proceselor biologice sub diferite forme. Sunt discutate subcapitole legate de modelarea matematică a proceselor biologice, continuând cu tehnici inteligente care emulează funcții specifice ființelor umane sau ale altor sisteme biologice (fuzzy, rețele neuronale, algoritmi genetici) și informații legate de alinieri ale expresiilor de gene. Capitolul se încheie cu descrierea unor multiple analize de expresii de gene, printre care amintim cele pe baza modelului probabilistic Bayes, al rețelelor bayesiene, al modelului Markov ascuns, metoda celor mai mici pătrate, analiza componentelor principale, sau al grupării ierarhice.

În Capitolul 4 este discutată pe larg analiza de microarray, și care sunt principalii pași într-o astfel de analiză genetică. Vorbim de partea de achiziție a datelor, partea de preprocesare și reducere a datelor, prin eliminarea atributelor redundante, continuând cu prezentarea mai multor metode de grupare și de investigare a expresiei diferențiale prin diferite teste sau metode statistice, cu specificarea tipurilor de modele potrivite analizei biologice și modalitățile de validare și evaluare a acestora.

Capitolul 5 prezintă contribuțiile științifice aduse în cadrul tezei. Începe cu prezentarea cadrului de implementare pentru realizarea obiectivelor propuse, continuând cu abordarea proprie privind principalii pași de parcurs pentru o analiză completă a expresiilor de gene. Astfel, partea I conține pașii de normalizare și reducere a dimensiunii datelor până la identificarea genelor exprimate diferențial, urmată de partea a II-a ce conține pașii de asociere a genelor

exprimate diferențial în module de corelație, identificarea genelor hub, precum și crearea de grafuri de co-expresie în funcție de relațiile dintre gene. Partea a III-a presupune o amplă caracterizare a genelor corelate cu cancerul de sân, identificate în urma adnotarii, urmată de partea a IV-a, o modelare de gene critice specifice cancerului, în termeni de rețele bayesiene, în vederea exemplificării și identificării unor posibile trasee de alterare a căilor celulare care pot duce la promovarea cancerului mamar.

Capitolul 6 prezintă rezultatele obținute în urma modelării rețelei genice selectate ca având cea mai mare relevanță biologică în cancerul de sân, analiza critică a rezultatelor și demonstrarea atingerii obiectivelor. Cu ajutorul inferenței bayesiene, au fost analizate efectele pe care posibilele mutații sau polimorfisme apărute în fiecare genă, modifică sau nu rolul genei ADAM28 din modul, aleasă ca fiind reprezentativă în progresia tumorală a cancerului de sân.

Capitolul 7 reprezintă sinteza realizărilor și contribuțiilor științifice aduse, punctând de asemenea și limitele cercetărilor întreprinse. Totodată, sunt indicate elementele de originalitate și perspective de dezvoltare viitoare. Capitolul se încheie cu prezentarea bilanțului activității de cercetare efectuate de-a lungul anilor de activitate doctorală.

Obiectivul principal al tezei de doctorat a fost de a identifica genele a căror expresie este corelată cu o anumită trăsătură fenotipică, în cazul nostru cancer de sân și care pot fi utilizate ca instrumente pentru realizarea unor posibile trasee prin care mutații ale genelor exprimate diferențiat să ducă la apariția acestei neoplazii. Abordarea descrisă în prezenta lucrare este utilă în descoperirea și înțelegerea cât mai completă a relațiilor cauzale dintre genele identificate ca find exprimate diferențial dintr-un set biologic de date.

Originalitatea acestei lucrări constă în analiza integrată a unor expresii de gene din țesut mamar tumoral. Analiza, realizată în limbajul R, începe printr-o dublă filtrare a setului de gene, cu ajutorul a două teste statistice: Chi-pătrat și Welch, la sfârșitul cărora a rezultat un număr redus de gene exprimate diferențial cu semnificație statistică. Pe baza acestor gene, cu ajutorul unei analize de corelație, au fost identificate module genice, din care, pe baza legăturii de corelație ale top 30 de gene din fiecare modul, au fost create rețele de co-expresie genică. Din totalul de rețele rezultate, a fost selectat modulul de co-expresie genică cu cea mai mare relevanță biologică (din toate cele identificate de noi) în cancerul de sân. Reproducerea schematică a modulului ales a fost efectuată cu ajutorul rețelelor bayesiene, în mediul de dezvoltare Netica. În scopul studierii și identificării principalelor relații ce stau la baza aparițiilor

mutațiilor în gene, diverse probabilități au fost atribuite variabilelor, supunând modelul la influențe diverse, cu scopul de a determina caracteristicile pe care sistemul le poate avea, înainte de a dezvolta cancer. Accentul a fost pus cu precădere pe modelarea rolului genei efect ADAM28 în cadrul rețelei, în scopul concluzionării asupra posibilelor motive de apariţie şi promovare a cancerului de sân, fiind cunoscut faptul că mutații în gena ADAM28 induc progresia cancerului prin diferite mecanisme.

Capitolul 2. Expresia genelor

2.1. Bioinformatică și data mining

Bioinformatica este un domeniu de cercetare foarte activ și atractiv, cu un impact ridicat în dezvoltarea noilor tehnologii, care combină domenii variate precum biologia, medicina, matematica sau știința computerelor [1, 2]. Este definită ca cercetare, dezvoltare și aplicare a instrumentelor de calcul pentru utilizarea datelor biologice, medicale, comportamentale sau de sănătate [2]. Din perspectiva tehnologiei informației, bioinformatica este o disciplină științifică care cuprinde achiziția, depozitarea, prelucrarea, analiza, interpretarea și vizualizarea informațiilor biologice [3].

2.2. Reglarea expresiei genelor

Materialul genetic este conținut într-o structură numită cromozom. Fiecare cromozom este alcătuit dintr-un lung șir de ADN asociat cu proteine, reprezentând o cromatină condensată, adică un complex de ADN și proteine. Funcția cea mai importantă a lor este aceea de conținere de gene. Mai exact, pentru om, există 23 de perechi de molecule de ADN, ce formează cromozomii [4]. Fiecare celulă umană are 46 cromozomi, ce se grupează în perechi, rezultând astfel 23 de perechi, din care, din fiecare pereche, un cromozom provine de la mamă, iar celălalt de la tată. Fiecare cromozom conține un centromer și doi telomeri, fiecare dintre aceștia

conţinand o „spirală”, fir unic continuu de ADN, ce are ca rol păstrarea și transmiterea corectă a informațiilor ereditare în cursul diviziunii celulare.

Gena este segmentul de ADN ce posedă instrucțiunile de formare a unei proteine. O genă reprezintă o secvență de nucleotide din molecula de ADN, folosită pentru producerea de proteine și ARN. Molecula de ADN are formă de spirală dublă, structura dublu catenară a ADN-ului (dublu helix) fiind asigurată de legăturile de hidrogen ce se stabilesc între bazele azotate ale celor două catene de ADN [5]. Nucleotidele dintr-un șir se leagă cu celelalte de pe al doilea șir astfel: adenina se leagă cu timina și invers, iar guanina se leagă cu citozina și invers. Cele două catene ADN sunt complementare și antiparalele, citirea lor făcându-se în direcția 5’ – 3’, de la capătul fosfat la hidroxil. În timpul diviziunii, ADN-ul se desface la mijloc, iar fiecare dintre cele două lanţuri de ADN este folosit ca matriţă pentru formarea părţii complementare în duplicarea ADN-ului. ”Treptele” moleculei originale de ADN, în formă de spirală dublă, se rup în locul legăturii de hidrogen şi astfel apar două jumătăţi de „treaptă” ale căror nucleotide se unesc cu cele libere, refăcându-se treptele moleculei, rezultând două molecule noi şi identice cu ADN-ul original, deoarece succesiunea bazelor este aceeaşi.

Bazele azotate sunt molecule care intră în alcătuirea nucleotidelor, ARN-ului și ADN-ului. Moleculele de ADN și ARN conțin cinci baze azotate: adenina (A), citozina (C), timina (T), guanina (G) și uracilul (U). Primele patru baze azotate (A, C, T și G) intră în compoziția ADN-ului, în timp ce în compoziția ARN-ului, timina (T) este înlocuită cu uracilul (U).

Înțelegerea structurii ADN-ului este primul pas în înțelegerea modului în care informația biologică este conținută în gene.

Procesul de reglare a expresiei genelor este exemplificat in Figura 2.1 si surprinde cadrul sintezei proteinelor, in care au loc două procese, unul de transcripție, reprezentând sinteza de molecule de ARN și unul de translație, la care ia parte doar ADN-ul codificator, adică genele, în care o moleculă de ARN sintetizează direct o proteină.

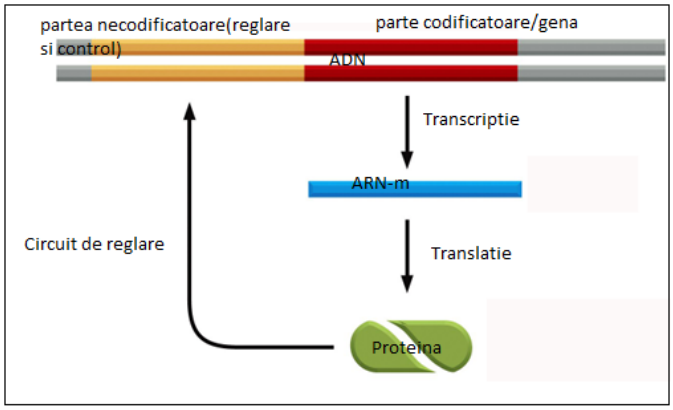

Figura 2.1. Biosinteza proteinelor [5]

În cadrul procesului de transcripţie ce are loc la nivelul nucleului, informaţia genetică din ADN se copiază în ARN-ul mesager în direcţia 5'-3', cu modificarea bazei azotate timina, care va fi înlocuită în ARN cu uracilul (T devine U) sau altfel spus transcripția are loc când porțiunea codificatoare a unei gene este „rescrisă" sub forma unei catene de ARN complementar numită ARN mesager (ARNm) și este efectuată de un complex proteic numit ARN polimerază care leagă regiunea promotor a genei și apoi „umblă" de-a lungul ADN-ului, catalizând formarea ARNm de la precursorii nucleotidici. Factorii de transcripție controlează reglarea genelor prin legarea la porțiuni specifice ale ADN-ului. Legarea factorului de transcripție poate duce la activarea sau reprimarea transcrierii, fie prin provocarea unor modificări structurale în ADN, fie prin interacțiuni cu proteinele care transcriu direct ADN-ul în ARNm. Factorii de transcripție ai siturilor de legare sunt secvențe relativ scurte de nucleotide (de obicei 5-15 nucleotide în lungime) [6].

La nivelul translaţiei ce are loc la nivelul ribozomilor, se produce decodificarea ARN-ului mesager, acest nivel fiind responsabil pentru sinteza proteinelor. Ribozomul, o structură foarte complexă, compusă din 2/3 ARN și 1/3 proteine, citește secvența nucleotidică și produce un lanț polipeptidic, astfel că în ribozomi vor ajunge grupări de câte trei nucleotide numite codoni, fiecărui codon corespunzându-i un anumit aminoacid (AA).

Cele două procese semnificativ diferite și complexe, implică fiecare un număr foarte mare de reacții biochimice (dintre care multe nu au fost pe deplin caracterizate). Alterări ale reacțiilor biochimice la aceste niveluri pot duce la dereglări în codul genetic și astfel la apariția diverselor boli, printre care și cancerul.

2.3. Tehnici de profilare a expresiei genelor

Pe măsură ce cunoștințele privind baza biologică și biofizică a funcției celulare au crescut, s-au extins oportunitățile pentru a avansa înțelegerea funcționării la scară celulară și moleculară a materiei organice, precum și pentru proiectarea aplicațiilor în diverse domenii de tratament medical și diagnostic.

Expresia genică bazată pe microarray a propulsat cunoștințele noastre în materie de biologie moleculară [7]. În primul rând, microarray-urile cu expresie genică au devenit o tehnică larg utilizată pentru a studia dinamica proceselor biologice, fiind niște laboratoare miniaturizate pentru studiul expresiei genelor [6]. Matricile de gene măsoară nivelul de expresie ARN molecular pentru mii de gene simultan, tehnica fiind o metodă de colectare a datelor foarte necesară pentru obținerea de informații legate de înțelegerea complexității organismelor vii [8]. Putem obține răspunsuri la componente biologice care interacționează unele cu altele, furnizându-ne informații legate de o înțelegere cât mai bună a sistemului biologic ales, analiza microarray având diverse aplicații în domeniul medical, pornind de la caracterizarea tumorilor benigne sau maligne, la evoluția sau modificările bolilor sau simptomelor în timp sau răspunsul la medicamente și identificarea de noi tratamente [9]. Transcriptomul celular normal poate fi comparat cu transcriptomul unei boli specifice, pentru a incerca elucidarea modificărilor specifice bolii. O altă aplicație poate fi analiza modificărilor fiziologice de-a lungul vieții, de exemplu comparația transcriptomului tânăr cu unul bătrân [10], dezvăluind modificări în căile moleculare.

Practic, un microarray ADN este o colecție de puncte microscopice atașate la o suprafață solidă, necesare pentru a măsura nivelurile de expresie ale genelor. Această tehnologie permite cercetătorilor să studieze simultan un număr mare de gene (aproximativ 21.000 de gene din genomul uman) [11]. Datele matricei de expresie genică pot fi analizate pe cel puțin trei niveluri de complexitate [9]. Primul nivel este cel al genelor individuale, unde se caută dacă o genă izolată se comportă diferit într-o situație de control versus o situație de tratament. Al doilea nivel este cel de gene multiple, în care grupuri de gene sunt analizate în termeni de funcționalități comune sau interacțiuni. Al treilea nivel încearcă să deducă gene și rețele de proteine care sunt responsabile de tiparele observate.

Tehnologiile de ADN microarray generează multe profile de expresii genice. În prezent, se folosesc două tehnologii de microarray și anume ADN complementar (ADNc) și oligonucleotide [12]. Amandoua presupun hibridizarea, ceea ce diferă fiind așezarea secvențelor de ADN precum și lungimea secvențelor.

Tehnicile ADNc au lungimi de baza de la câteva sute la câteva mii de probe. De obicei, pentru majoritatea experimentelor de profilare a expresiei genelor cu microarray ADNc, ARNm din două surse diferite (precum ar fi celulele bolnave și celulele normale) este extras, purificat și făcută operația de transcripție inversă în prima catena de secvențe ADNc. Fiecare lot de ADNc este marcat cu un fluorofor diferit. După marcare, probele de ADNc sunt amestecate și hibridizate cu același microarray. După hibridizare, un microscop laser este folosit pentru scanarea petelor. Emisiile individuale din coloranți la fiecare punct sunt înregistrate și stocate. Secvențele de ADN sunt marcate prin diferite culori (cele mai folosite fiind verde și roșu) și situate în diferite puncte pe suprafață de microarray, indicând niveluri mai mari sau mai mici de cantități (expresii) de gene din acea mostră.

Matricile de tip oligonucleotidice presupun folosirea unui chip numit Affymetrix GeneChip, unde expresia fiecărei gene este măsurată prin compararea mostrei de ARNm hibridizat cu un set de probe compuse din 11-20 perechi de oligonucleotide, fiecare având lungimea de 20-30 de nucleotide (perechi de baze). Primul tip de probă din fiecare pereche poartă numele de perfect match (PM) și este luată din secvență de gene. Al doilea tip de probă se numește mismatch (MM) și se crează schimbând cea de-a 13-a genă din PM pentru a reduce rata îmbinării specifice ARNm-ului pentru acea genă. Pentru fiecare genă (probe set) se obțin doi vectori de intensități, unul pentru PM, altul pentru MM.

Diferența dintre cele două canale este dată de hibridizarea diferită: ARNm la microarray-uri de tip oligonucleotidice, față de ADNc la microarray-uri pe care sunt hibridizate două tipuri de celule. De asemenea, tehnica ADNc este o tehnică pe două canale, iar tehnica ARN pe un singur canal, tipurile de probe fiind sintetizate direct pe microarray.

Comparativ cu alte instrumente de biologie, microarray-urile genomice reprezintă platforme care permit accesul mai facil la mecanismele biologice interne ale culturilor celulare. Însă, în timp ce seturile mari de date generate de microarray-uri reprezintă o potențială mină de aur a informațiilor biologice, dimensiunea lor este cea care face procesarea datelor o sarcină greoaie. Acest lucru se poate complica și mai mult prin efectele inevitabile de lot generate atunci

când sunt combinate diferite seturi de date. Mai mult, profilurile de expresii de gene sunt dependente de combinații de evenimente intracelulare complexe și, ca atare, identificarea semnalelor legate în primul rând de fenotipul de interes reprezintă o provocare substanțială.

Capitolul 3. Prezentarea sintetică a contribuțiilor științifice

3.1. Cadrul de implementare

Datele provenite din expresiile genelor pot conduce la aplicații complexe precum descoperirea unor noi gene, diagnosticarea diferitelor boli, descoperirea de medicamente sau cercetarea toxicologică.

Cu o cantitate atât de mare de date disponibile publicului larg, este esențial ca un analist bioinformatic să posede cunoștințele și abilitățile specifice pentru a înțelege, analiza și interpreta aceste date într-un mod cât mai corect.

În această lucrare, vor fi analizate date genice provenite din matrici oligonucleotidice, de pe chipuri numite Affymetrix GeneChip, în limbajul de programare R, un mediu special folosit pentru manipularea datelor, calcul statistic și afișare grafică, care oferă o gamă largă de tehnici statistice (modelare liniară, neliniară, teste, analiza seriilor de timp, clasificare, grupare), inclusiv o colecție vastă de instrumente de analiză, manipulare și stocare a datelor [13]. În matricile oligonucleotidice, expresia fiecărei gene este măsurată prin compararea mostrei de ARNm hibridizat cu un set de probe compuse din 11-20 perechi de oligonucleotide, fiecare având lungimea de 20-30 de nucleotide (perechi de baze), fiecare genă (probe set) fiind reprezentată cu ajutorul a doi vectori de intensități, unul pentru PM, altul pentru MM.

Capitolul 3 prezinta principalii pași parcurși pentru analiza biologică a expresilor genice, printre care amintim: descărcarea fișierelor CEL, încărcarea și normalizarea datelor, filtrarea setului de date, găsirea genelor exprimate diferențial, gruparea genelor în clustere și analiza lor, adnotarea la simboluri genice și găsirea de relevanțe biologice în conformitate cu tema aleasă.

Partea I conține pașii de analiză microarray realizați în limbajul R până la identificarea genelor exprimate diferențial, văzute ca potențiali markeri tumorali (citirea expresiilor de date,

normalizarea, filtrarea, reducerea dimensiunilor), urmată de partea a II-a ce conține pașii de analiză și împărțire a genelor exprimate diferențial în module de corelație, identificarea genelor hub, precum și crearea de grafuri de co-expresie în funcție de relațiile dintre gene. Partea a III-a presupune o amplă analiză a genelor corelate cu cancerul de sân, urmată de partea a IV-a, o modelare de gene critice specifice cancerului, în vederea exemplificării și identificării unor posibile trasee de alterare a căilor celulare care pot duce la promovarea cancerului mamar.

3.2. Metodologie

În această lucrare, am utilizat pentru analiza expresiilor genice ale cancerului de sân, datele de expresie ale fișierului GSE48391 din baza de date NCBI Gene Expression Omnibus [14], fișier din care au fost selectate doar datele de expresie genică din țesut de sân canceros. Seria GSE conține liste de fișiere GSM ale fișierelor Affymetrix CEL, care formează împreună un singur experiment, în care fișierele GSM reprezintă date la nivel de probă din utilizarea unui singur cip, sub forma intensităților PM și MM. În seria GSE amintită, pentru fiecare dintre probe/pacienți, există un set de caracteristici cu valorile lor specifice de activitate la un anumit moment de timp, într-un anumit stadiu al cancerului de sân, reprezentate sub formă de matrici oligonucleotidice provenite de pe un chip numit Affymetrix Human Genome U133 Plus 2.0 Array [2HG-U133_Plus_2].

Toate cele 81 de fișiere ce conțin date referitoare la cancerul de sân au fost descărcate și analizate cu scopul analizei principalelor gene care scapă din echilibru și conduc spre multiplicare infinită, necontrolată, reușind să treacă cu success de barierele corpului.

Înainte de a începe efectiv analiza datelelor, au fost efectuate reprezentări ale datelor înainte de normalizare, cu datele brute, trasând atât un boxplot cu valori de intensitate nenormalizate, cât și histograma densitate vs log intensitate pentru datele nenormalizate, pentru a vedea după normalizare, efectele acestei tehnici asupra expresiilor de gene.

Normalizarea datelor microarray este necesară pentru a se asigura că diferențele de intensitate citite de scaner se datorează expresiilor genice diferențiate și nu datorită imprimării, hibridizării sau artefactelor de scanare, înțelegând ajustările realizate pentru erorile sistematice introduse de diferențele de proceduri și efectele de intensitate ale coloranților etc.

Algoritmul Robust Multiarray Averaging (RMA) din pachetul affy va fi utilizat în această lucrare, pentru a normaliza toate fișierele CEL. Constând din trei pași și anume: corecția de fundal a valorilor PM pe fiecare matrice separat, normalizarea și rezumarea măsurii expresiei genice, analiza RMA folosește doar informațiile din probele PM pentru a estima parametrii de distribuție și a întoarce semnalul estimat. Funcția RMA [15] setează cipurile la aceeași distribuție și aceeași medie și calculează logaritmul din PM-uri, convertind obiectele de tip AffyBatch în obiecte de tip ExpressionSet, astfel având loc o preprocesare a datelor brute (o rezumare a măsurătorilor probelor într-o singură măsurare per probă). În interiorul unei matrici ExpressionSet se găsesc nivelurile de expresie genică ale genelor i în proba de ARNm j, reprezentate prin valori normalizate log2. După normalizare, putem verifica efectele RMA printr-un boxplot cu valori de intensitate normalizate.

Dimensiunea datelor normalizate este de 54675 (număr de gene) x 81 (număr de probe), astfel că se ridică problema dimensionalității ridicate și în acest caz, un pas important ce trebuie să fie făcut fiind cel de reducere a dimensiunilor.

În cadrul acestei analize, am selectat genele care sunt exprimate în cel puțin 5% din eșantioane, care au o varianță semnificativ diferită (mai mare) de varianța mediană a tuturor seturilor de probe. Astfel, din matricea de expresie a setului de date, pentru fiecare genă în parte, am selectat doar valorile expresiilor genelor mai mari de a 5-a decilă, care să fie exprimată în cel puțin 5% din numărul total de probe existente în matrice. Apoi cu ajutorul unui test Chi-pătrat și folosind un prag ales (valoarea $p < 0.01$), am aplicat un test qchisq pentru alegerea genelor cu o varianță mai mare decât cea medie. Testul Chi-pătrat pentru varianță este o procedură non-parametrică în care se folosește o statistică distribuită de test Chi-pătrat, utilizată pentru a determina dacă variant unei variabile este diferită față de o altă valoare specificată, adică dacă există sau nu o legătură între valoarea rezultată și o valoare așteptată.

Folosind funcția cuantilă qchisq() din R, a fost calculată distribuția actuală Chi-pătrat, care apoi va fi comparată cu testul statistic Chi-square pentru varianță. Astfel, se va decide dacă varianța unei gene este sau nu diferită față de cea calculată cu ajutorul unui prag ales și a gradelor de libertate. Dacă valoarea din testul statistic este mai mare decât valoarea rezultată în urma aplicării funcției qchisq [16], atunci genele vor fi păstrate, altfel nu.

Livrabilele vor fi un număr mult mai mic de gene, gene care au șansa să caracterizeze un fenotip care să se diferențieze de restul populației, noua matrice având dimensiuni mult mai

reduse față de cea inițială, în urma acestor pași de filtrare. În cazul de față, reducerea dimensiunii datelor după cele două etape, ajunge la un număr de 19847 (gene) la cele 81 de probe existente.

Pentru etapa de selectare a genelor exprimate diferențial, vor fi identificate și selectate genele exprimate diferențial, cu o valoare p semnificativ statistică, cu scopul de a determina dacă există subgrupuri de pacienți cu profiluri diferite (markeri moleculari diferiți).

Ulterior, dendograma rezultată ca urmare a grupării ierarhice folosind funcția hclust, a fost tăiată într-un număr de trei clustere, pe baza acestora identificând genele exprimate diferențial între un cluster dat și toate celelalte, folosind un test t tip Welch, conceput pentru variații inegale ale eșantioanelor și/sau dimensiuni diferite ale probelor, dar cu păstrarea presupunerii distribuției normale. Altfel spus, am testat dacă media primului cluster este egală sau diferită față de mediile celorlalte clustere, caz în care se aplică un test "cu o singură coadă", o modalitate alternativă de calcul a semnificației statistice, utilizată pentru distribuții asimetrice, care calculează și valoarea nivelul de semnificație statistică p [17]. Cu cât valoarea p este mai mică, cu atât sunt mai puternice dovezile că ipoteza nulă ar trebui respinsă, astfel că o valoare p mai mică de 0.05 este semnificativă statistic. Pe baza acestui criteriu, am selectat lista genelor exprimate semnificativ diferențial, stabilind o valoare prag <0.05, rezultând un număr de 7189 gene x 81 eșantioane.

În continuare, pe baza matricii nou rezultate, am folosit pachetul Weighted Correlation Network Analysis (WGCNA) din limbajul R [18], care formează grupuri de gene corelate între ele, rezultând module relaționale în care sunt identificate și selectate genele "conducătoare"/hub-uri intramodulare din modulele de consens, definite ca gene cu cea mai mare corelație din modulul respectiv.

Sunt construite două matrici, mai întâi o matrice de adiacență, metoda de corelație standard fiind corelația Pearson la o puterea de prag aleasă conform unui criteriu de selecție, cu informații despre corelația dintre valorile expresiilor dintre gene, urmată de definirea unei matrici TOM (Topology Overlap Matrix), de "vecinătate", care ia în considerare asemănarea topologică și similaritățile dintre gene, reflectate la nivel de topologie de rețea [19].

Ulterior, este realizată o noua grupare ierarhică pentru împărțirea genelor care tind să aibă o conectivitate ridicată, în module de co-expresie. Modulele genice reprezintă grupuri de gene puternic interconectate în termeni de co-expresie, identificarea acestora putând fi făcută cu

ajutorul culorilor, pentru o mai ușoară gestionare a lor. În Figura 3.1 sunt prezentate modulele genelor corelate în corespondență cu harta grupată a genelor.

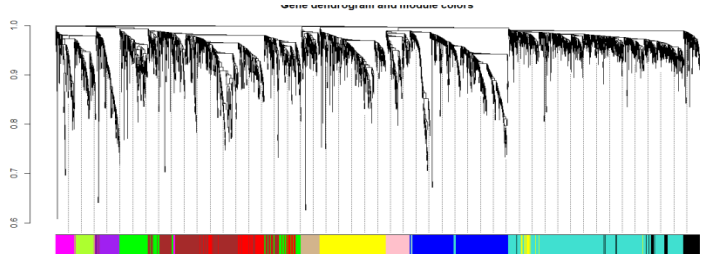

Figura 3.1. Guparea genelor în funcție de matricea TOM și modulele rezultate

Numărul minim de gene per modul a fost ales 100, rezultând astfel 13 module, dintre care 12 module cu gene corelate între ele, însumând un total de 3663 gene și un modul în care se găsesc gene necorelate funcție de distanță.

În Figura 3.2, pot fi văzute atât modulele de gene co-exprimate, cât și modul de grupare și relațiile dintre module. Practic, harta grupată a genelor din Figura 3.1 este mapata în arborele din figura de mai jos, a celor 12 module de co-expresie genică.

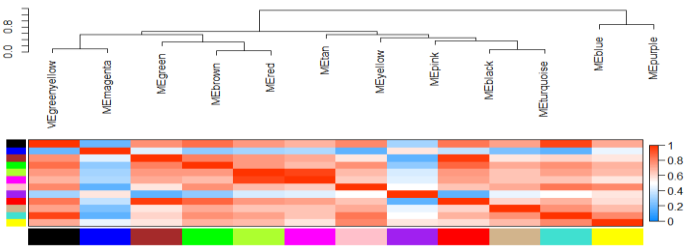

Figura 3.2. Modulele de co-expresie genică grupate în funcție de distanța euclidiană

Pentru fiecare modul, au fost calculate matricile de adiacență pentru genele în cauză, iar pentru verificarea și validarea modulelor atribuite, am optat și pentru aplicarea funcției eigengen, care calculează PCA pentru expresiile de date, caz în care modulele similare sunt asimilate. Modulul eigengen este definit ca prima componentă principală a unui modul dat, sub forma unui vector de date unidimensional, considerat un reprezentant al profilurilor/datelor de expresie genică din acel modul.

Următorul pas în analiza genelor a fost identificarea acelor gene din fiecare modul, care sunt cel mai conectate în cadrul modulului (conectivitate intramodul ridicată), adică acele gene hub cel mai puternic corelate cu o trăsătură clinică sau fenotipică de interes (în cazul nostru cancerul de sân).

Pentru a afla genele (hub-urile din fiecare modul) cu cele mai mari conectivități intramodulare, astfel încât suma greutăților marginilor din modul să fie maximă, din fiecare matrice de adiacență (fiind matrice inversabilă), au fost calculate maximele pe coloana/linie pentru fiecare modul în parte [20]. Astfel, au fost alese top 30 gene din fiecare modul, în funcție de valorile descrescătoare ale rezultatelor, gene pentru care au fost create noi matrici de adiacență.

Fiecare genă hub rezultată în fiecare modul a fost adnotată cu ajutorul pachetelor "hgu133plus2.db", AnnotationDbi.db" și "annotate", pentru fiecare id de genă fiind identificate numele/simbolul afferent și Entrez id-ul, un identificator unic pentru fiecare genă, utile pentru căutarea de informații în baze de date cu conținut genomic, precum determinarea sau stabilirea proceselor biologice cu care acestea au legătură.

Folosind pachetul igraph din limbajul de programare R, au fost citite matricile de adiacență pentru fiecare top 30 gene din fiecare modul și au fost create fișiere edge list, reprezentând corelațiile celor 30 gene din fiecare modul împreună cu greutatea (weight) sa, de forma "from to weight" și astfel generate scoruri de corelație între noduri (genele hub).

Cu ajutorul acestor fișiere, au putut fi create grafuri sau rețele de co-expresie genică pentru fiecare din cele 30 de gene corelate din fiecare modul. O rețea de co-expresie genică (GCN) este un grafic nedirecționat, în care fiecare nod corespunde unei gene și o pereche de noduri este conectată cu o margine dacă există o relație semnificativă de co-expresie între ele.

Pentru fiecare culoare/modul am sortat genele în funcție de valoarea weight, selectând primele 30 de corelații în ordinea valorilor descrescătoare a weightului. Entrez_ID-urile și simbolurile genice au fost utilizate pentru a găsi informații despre fiecare genă din fiecare modul, ca având sau nu o legătură cu o trăsătură fenotipică de interes, marker, promotor sau genă specifică cancerului de sân.

Fiecare genă/proteină de mai sus a fost căutată și validată în "The Human Protein Atlas" [21], ca fiind detectată în unul sau mai multe țesuturi din corpul omenesc, cu un grad mai mare sau mai mic de apariție în diverse tipuri de cancer. Din aceste gene, am selectat un număr de zece

gene cu un prognostic favorabil sau nefavorabil în apariția cancerului de sân [21], și anume cinci gene identificate din modulul grey, două din modulul purple și câte una din modulul greenyellow, turquoise și yellow. Pentru fiecare modul în parte, au fost extrase informații și caracteristici despre fiecare genă în parte.

Potrivit teoriei funcției hclust [22], grupul cu genele cele mai corelate se află în partea stânga a dendogramei, în cazul nostru modulul greenyellow conținând genele cel mai puternic corelate, astfel că în continuare, a fost ales modulul "greenyellow" pentru o analiză mai detaliată a genelor și a legăturilor dintre acestea, în lupta cu cancerul de san. Astfel, în afară de RAC2, au fost aduse informații biologice și pentru genele: ADAM28, ARHGAP9, STK17A, CLEC12A, CSF2RB, LY86, TASL, SELL, cu rol important în progresia celulelor cancerului mamar în organismul uman.

Figura 3.3 prezintă rețeaua de co-expresie genică (GCN) a modulului cu genele cele mai corelate, conform criteriilor de selecție ale pachetului de analiză a genelor de co-expresie, graf în care fiecare nod corespunde unei gene, fiecare pereche de noduri fiind conectată cu o margine dacă între noduri (gene) există o relație semnificativă de co-expresie.

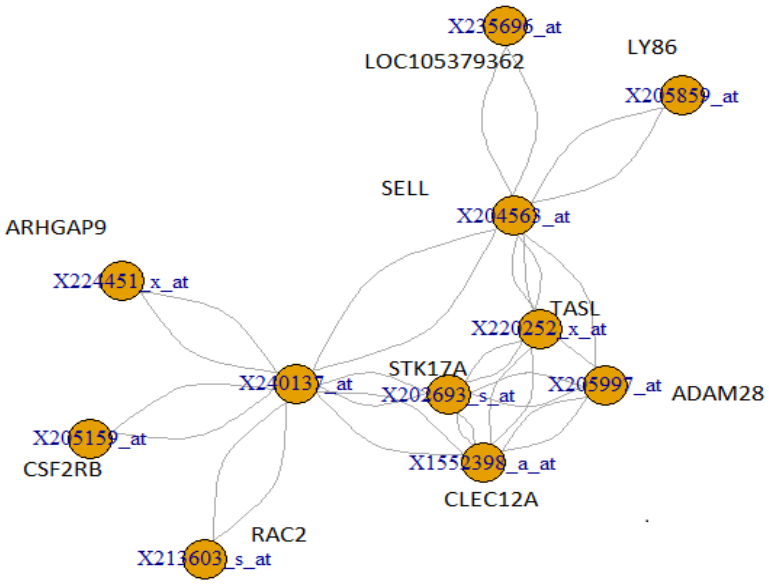

Figura 3.3. Rețeaua de co-expresie genică a modulului greenyellow

După cum se poate observa, graful este nedirecționat, sensul legăturilor dintre gene fiind bidirecțional. Pentru exemplificarea posibilelor căi de alterare care duc la formarea celulelor

canceroase, pornind de la genele analizate în partea a III-a, a fost aleasă calea unidirecțională din Figura 3.3.

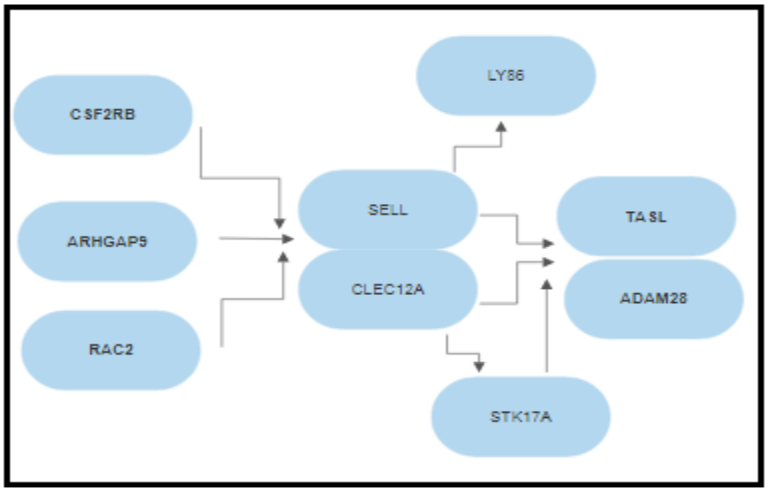

Figura 3.4. Reproducere schematică a sistemului genic din cadrul modulului ales a fi modelat

Conform reprezentării schematice de mai sus (Figura 3.4), rețeaua modelată pornește de la genele CSF2RB, RAC2 și ARHGAP9 până la TASL și ADAM28. Ideea principală a acestei modelări a constat în găsirea unor posibile trasee ale mutațiilor genelor amintite, pentru a putea da naștere la o proliferare celulară necontrolată și ulterior, la o invazie în celulele vecine.

Obiectivul principal al tezei de doctorat a fost cel de realizare a unor trasee prin care eventualele mutații ce pot avea loc în interiorul genelor exprimate diferențiat, identificate în etapele descrise anterior și analizate amănunțit cu ajutorul bazelor și bibliotecilor de date publice, să ducă la apariția cancerului de sân. Astfel, cu ajutorul inferenței bayesiene, au fost analizate efectele pe care posibilele mutații sau polimorfisme apărute în fiecare genă, modifică sau nu rolul genei ADAM28 din modul, aleasă ca fiind reprezentativă în progresia tumorală a cancerului de sân.

Inferența Bayesiană este cunoscută drept una dintre cele mai bune abordări pentru modelarea incertitudinii, multe aplicații ce se bazează pe rețele bayesiene fiind create pentru modelarea oricărei variabile aleatoare, inclusiv indicatori de performanță în afaceri, inginerie, medicină sau ecologie [23]. Se bazează pe teorema lui Bayes, ce poate fi interpretată ca probabilitatea unor seturi de atribute de a aparține unei anumite clase:

$$P(H_i|S = s_1, \dots s_n) = \frac{P(S=s_1,\dots s_n|H_i)P(H_i)}{P(S=s_1,\dots s_n)} = \frac{P(S=s_1,\dots s_n|H_i)P(H_i)}{\sum_j P(H_j)P(S=s_1,\dots s_n|H_j)}, \quad (3.1)$$

unde $H = \{H_1, \dots, H_n\}$ reprezintă clasele, iar S atributele de clasificat.

Rețelele bayesiene reprezintă o clasă a modelelor probabilistice utilizate pentru modelarea raționamentului în condiții de incertitudine. Crearea unei rețele bayesiene presupune în principal parcurgerea a doi paşi:

- ➔ Construirea componentei calitative, adică graful aciclic direcţionat, conţinând noduri şi arce direcţionate (fapt ce presupune analiza problemei şi deducerea relaţiilor cauzale între variabile);
- ➔ Definirea componentei cantitative, presupunând tabelele de probabilităţi condiţionate ataşate fiecărui nod al reţelei şi care descriu incertitudinea asupra relaţiilor de (in)dependenţă dintre nodul respectiv şi părinţii săi direcţi.

Odată specificate probabilitățile condiționate ale variabilelor pentru toate combinațiile posibile ale părinților, rețelele bayesiene prezintă proprietatea că vor defini unic o factorizare a distribuției de probabilitate compusă peste setul variabilelor domeniului:

$$P(X_1, \dots, X_n) = P(X_1)P(X_2|parinte2) \dots P(X_n|parinten) = \prod_{i=1}^{n} P(X_i|X_{PA_i}) \quad (3.2)$$

Netica este un important software de dezvoltare a rețelelor Bayesiene, conceput pentru a fi performant și ușor de folosit, un instrument de analiză ales de multe dintre companiile și agențiile guvernamentale de top din lume [24]. Software-ul este un mediu de simulare pentru construirea şi evaluarea de reţele bayesiene, având avantajul unei interfeţe intuitive, care oferă suport de calcul pentru realizarea mai multor tipuri de inferenţă [25, 26]. Cu ajutorul Netica a fost creat un sistem direcționat de gene, format din noduri și arce, în care nodurile reprezintă genele, iar arcele legăturile de dependență dintre gene.

În cazul nostru, rețeaua bayesiană a fost creată exclusiv din gene susceptibile în cancerul de sân, identificate prin analiza R și validate de baze de date ce conțin informații despre cancerele umane [27]. În comparație cu diagrama din Figura 3.4, a fost eliminată gena SELL cu descendenții aferenți, restul de gene fiind implicate în mod activ (împreună sau separat) în

neoplaziile țesutului mamar, conform cercetărilor și studiilor efectuate până în present (Figura 3.5).

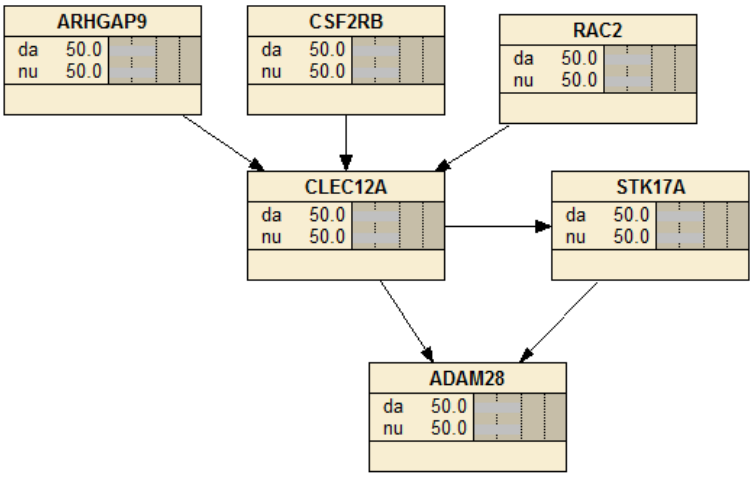

Figura 3.5. Rețeaua bayesiană aleasă a fi analizată

Din componenta calitativă a rețelei bayesiene fac parte toate cele șase noduri și relațiile de (in)dependență dintre ele, întâlnind în acest graf toate cele trei tipuri de conexiuni: seriale, convergente și divergente. Fiecare genă reprezintă, de fapt, un nod în rețeaua bayesiană creată, astfel că pentru a modela rețeaua avem nevoie de componenta cantitativă, adică atât de probabilități marginale ale variabilelor ataşate nodurilor, cât şi de tabele de probabilitate condiţionată asociate nodurilor, conform condiției de cauzalitate Markov [28].

În primă faza, am presupus fiecare genă ca având probabilități marginale egale, după care pentru probabilități diferite am modelat rețeaua pentru a analiza care sunt principalele mutații din gene care pot duce la transformări în structura genei ADAM28 sau altfel spus care pot fi cauzele mutațiilor din gena ADAM28, făcând referire la toate celelalte gene din proces.

Pentru cazul nostru, distribuţia de probabilitate compusă are forma:

$$P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)$$
$$= P(ARHGAP9)P(CSF2RB)P(RAC2)P(CLEC12A|ARHGAP9, CSF2RB, RAC2)$$
$$P(STK17A|CLEC12A)P(ADAM28|CLEC12A, STK17A) \quad (3.3)$$

iar probabilitatea mutațiilor apărute în gena ADAM28, dându-se celelalte gene, poate fi scrisă matematic folosind condiția de cauzalitate Markov, rezultând faptul că genele ADAM28 și RAC2, CSF2RB, ARHGAP9 sunt condițional independente:

$$P(ADAM28|ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A)$$
$$= \frac{P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)}{\sum_{ADAM28} P(ARHGAP9, CSF2RB, RAC2, CLEC12A, STK17A, ADAM28)}$$
$$= P(ADAM28 \mid CLEC12A, STK17A) \tag{3.4}$$

În continuare, vor fi prezentate şi analizate influenţele evidenţelor certe şi incerte asupra diferitelor noduri, pentru toate tipurile de conexiuni întâlnite în rețeaua bayesiană.

Astfel, am presupus pe rând informații certe asupra mutațiilor în gena CLEC12A, informații certe asupra mutațiilor în gena STK17A, urmate de cazuri cu informații certe asupra mutațiilor în ambele gene, pentru a vedea care sunt relațiile dintre gene și ce relații de (in)dependență condițională există între gena efect ADAM28 și nodurile cauză.

Prima presupunere a fost cea de informație certă în gena CLEC12A, astfel că probabilitatea ca gena ADAM28 să fie mutantă, știind gena CLEC12A mutantă este:

$$P(ADAM28|CLEC12A = da)$$
$$= \sum_{STK17A} P(STK17A|CLEC12A = da) P(ADAM28 \mid STK17A, CLEC12A = da) \tag{3.5}$$

Din compilarea rețelei multiplu-conectate (Figura 3.6), am putut observa faptul că:

- ➔ probabilitea marginală modificată la funcția de evidență peste distribuția de probabilitate (100 0) a genei CLEC12A, duce la modificări ale probabilităților pentru genele „parinți" RAC2, CSF2RB, ARHGAP9, dar și pentru gena „copil" STK17A ;
- ➔ probabilitatea marginală a genei ADAM28 se modifică la funcția de evidență peste distribuția de probabilitate (100 0) a genei CLEC12A;

- ➔ orice informație certă adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9, știindu-se funcția de evidență peste distribuția de probabilitate (100 0) a genei CLEC12A, nu modifică probabilitatea marginală a genei ADAM28, astfel că putem afirmă faptul că dându-se certă informația legată de gena CLEC12A, genele inițiale și gena finală sunt condițional independente.
- ➔ orice informație certă adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9, dându-se certă informația asupra genei CLEC12, nu modifică probabilitatea marginală a genei STK17A (genele inițiale și gena STK17A sunt condițional independente).

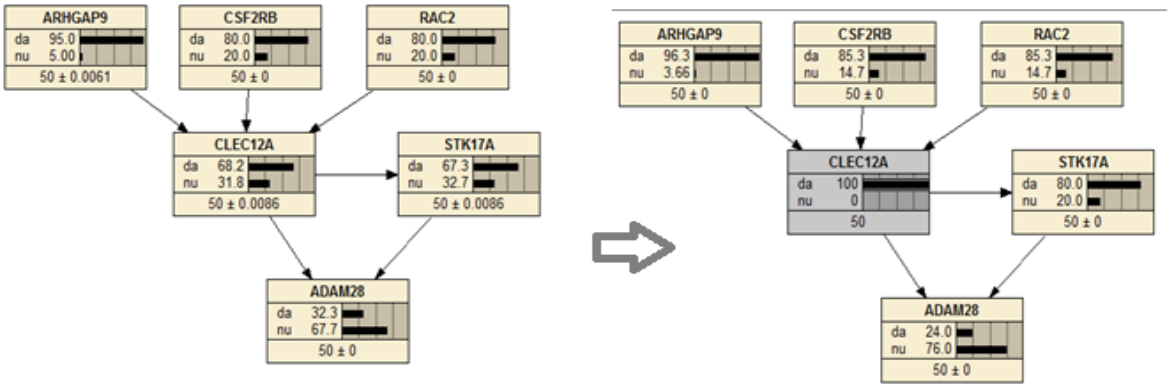

Figura 3.6. Modelarea rețelei la informație certă în gena CLEC12A

A doua presupunere a fost cea de informație certă în gena STK17A, astfel că probabilitatea ca gena ADAM28 să fie mutantă, știind gena STK17A mutantă este dată de formula:

$$P(ADAM28|STK17A = da\ ) = \sum_{CLEC12A} \frac{P(CLEC12A|ARHGAP9, CSF2RB, RAC2\ )}{P(ADAM28\ |\ CLEC12A, STK17A = da)} \quad (3.6)$$

Din compilarea rețelei (Figura 3.7), am putut observa faptul că:

- ➔ probabilitea marginală modificată la funcția de evidență peste distribuția de probabilitate (100 0) a genei STK17A, duce la modificări ale probabilității genei „parinte” CLEC12A, dar și ale probabilităților marginale ale genelor ARHGAP9, CSF2RB, RAC2;

- ➔ probabilitatea marginală a genei ADAM28 se modifică la funcția de evidență peste distribuția de probabilitate (100 0) a genei STK17A;
- ➔ orice informație certă adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP, știindu-se funcția de evidență peste distribuția de probabilitate (100 0) a genei STK17A, modifică probabilitatea marginală a genei CLEC12A, și prin urmare a genei ADAM28;

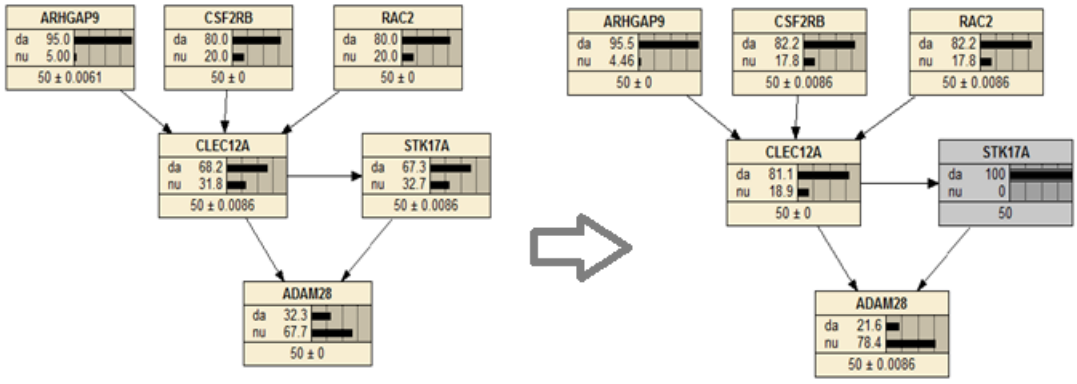

Figura 3.7. Modelarea rețelei la informație certă în gena STK17A

Cea de-a treia presupunere a fost cea de informație certă în ambele gene STK17A și CLEC12A, astfel că probabilitatea ca gena ADAM28 să fie mutantă, știind că cele două gene sunt mutante este dată de datele din tabelul de probabilități condiționate aferente genei ADAM28, cazul în care $STK17A = da, CLEC12A = da$.

Din compilarea rețelei (Figura 5.8), am putut observa faptul că:

- ➔ probabilitățile marginale modificate la funcția de evidență peste distribuția de probabilitate (100 0) a genelor STK17A și CLEC12A, duce la modificări ale probabilităților marginale ale genelor ARHGAP9, CSF2RB, RAC2 (modificările apar imediat după informația certă a genei CLEC12A, apariția unei informații certe și pentru gena STK17A nemodificând încă o dată probabilitățile rezultate;
- ➔ probabilitatea marginală a genei ADAM28 se modifică la funcția de evidență peste distribuția de probabilitate (100 0) a genelor STK17A și CLEC12A și depinde doar de linia din tab1elul de probabilități condiționate aferent genei ADAM28 $[P(ADAM28 \mid STK17\text{A} = da, CLEC12A = da)]$.

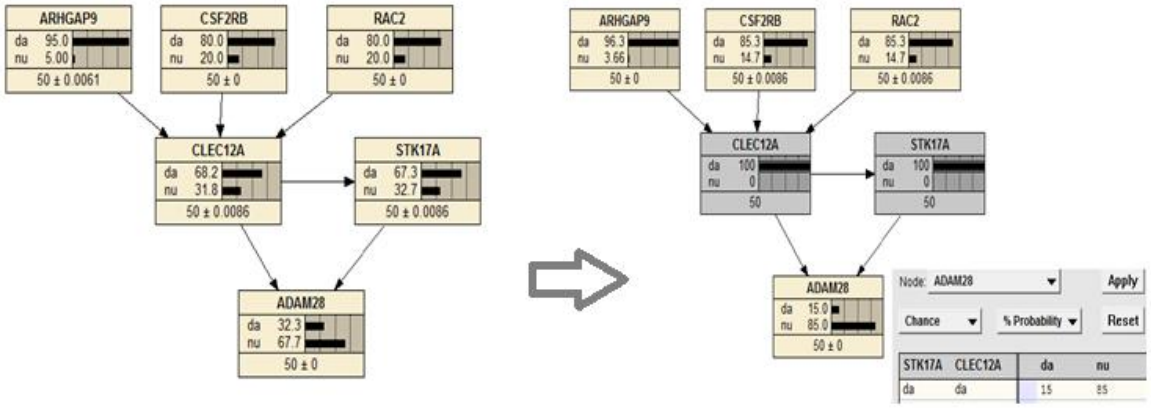

Figura 3.8. Modelarea rețelei la informație certă în gena STK17A și CLEC12A

Capitolul 4. Elemente de verificare și validare. Analiza critică a rezultatelor

Originalitatea acestei lucrări constă în modelarea unei rețele de co-expresie genică, cu relevanță biologică în cancerul de sân, modul genic ales după o analiză integrată în limbajul R a expresiilor de gene din fișierul GSE48391, bazată pe o abordare diferită față de celelalte existente. Pentru decoperirea și validarea unor biomarkeri moleculari pentru cancerul de sân am redus numărul de gene printr-o dublă filtrare până la identificarea acelor gene exprimate diferențial cu o valoare prag care să ateste o semnificație statistică. Pachetul WGCNA a fost utilizat pentru a identifica modulele și genele cheie reprezentative pentru studiul nostru. Pe baza corelațiilor și a asemănărilor topologice dintre gene, au fost construite rețele de co-expresie genică cu scopul de a identifica biomarkeri (gene hub) asociați cu progresia cancerului de sân.

În această lucrare, din matricea de expresie normalizată a setului de date GSE48391, am selectat genele exprimate în cel puțin 5% din eșantioane cu intensitatea semnalelor mai mare decât a 5-a decilă a valorilor expresiei genelor, cu o varianță mai mare decât cea mediană, folosind un prag $p < 0.01$ (test Chi-patrat). Apoi, folosind un test-t de tip Welch, pe baza unei grupări de probe într-un număr de trei clustere, folosind distanța euclidiană, au fost determinate pe baza unei valori $p < 0.05$, genele exprimate diferențial. În cazul nostru, reducerea dimensiunii datelor după cele două etape, ajunge de la un număr de 54675 de gene, la 19847 de gene la sfârșitul primei filtrări și la 7189 de gene la sfârșitul celei de-a doua. Cele 7189 de gene au fost

împărțite în 13 module de corelație, folosind logica WGCNA, dintre care 12 module cu gene corelate între ele și un modul în care se găsesc gene necorelate funcție de distanță.

Primul pas pentru verificarea și validarea modulelor atribuite a fost făcută aplicând funcția eigengen, care a calculat PCA pentru expresiile de date. Modulul eigengen este definit ca prima componentă principală a unui modul dat, sub forma unui vector de date unidimensional, și considerat un reprezentant al profilurilor/datelor de expresie genică din acel modul. În Figura 4.1 se poate observa validarea celor 12 module identificate în Capitolul 3, diferența dintre gruparea dinamică și funcția eigengen fiind asimilările modulelor similare ale celei din urmă (de exemplu, modulul magenta este asimilat de modulul greenyellow, cel red de modulul brown, iar cel turquoise asimilat de modulul black).

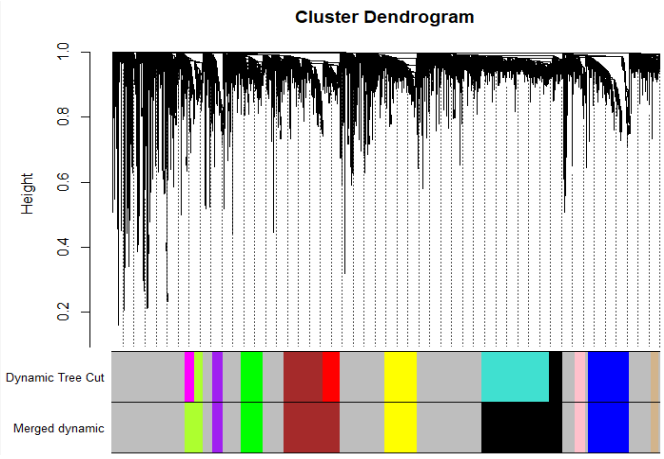

Figura 4.1. Validarea modulelor atribuite cu ajutorul funcției eigengen

Pentru validarea genelor din analiza noastră (GSE48391), au fost folosite și alte seturi de date independente. Pentru acest lucru, am descărcat fișierul GSE36295 [29], ce conține ARN izolat din țesuturi de cancer de sân excizate chirurgical, purificat, etichetat și hibridizat pe platforma Affymetrix Human Gene 1.0 ST Array. Pentru fișierul GSE36295, numărul de gene s-a redus de la 32321 x 50, la 8594 x 50 și apoi la 3685 x 50. De la 32321 gene și 50 probe, după filtrări am ajuns la 3685 de gene, din care au fost selectate un număr de gene din modulul cu genele cele mai corelate: SKAP2, ITGB5, MYCT1, PODNL1, TEK, SDR42E1, TFPI, JPH1, PAPSS2, MANEAL, PPIC, COBLL1, EMCN, toți markeri tumorali în diverse tipuri de cancer, însă gena de interes ADAM28 nu a putut fi găsită în matricea expresiilor de gene normalizate.

Un alt fișier GSE102907 [30], a fost descărcat, fișier ce conține ARN-ul mesager extras din tumora primară a unor pacienți cu cancer de sân, hibridizat și scanat cu matricea Affymetrix Human Genome GeneChip U133 Plus 2.0. De la 54675 gene și 61 probe, după filtrări am ajuns la 4541 de gene, din care au fost selectate un număr de gene din modulul ce conține genele cele mai corelate: MIPOL1, CDC20B, C7orf57, MPV17L, ZBTB18, CYP4F8, MYBPC1, KITLG, FAM110C, CSTF3, CROT, ARMC3, gena ADAM28 fiind găsită ca gena exprimată diferențial.

Cunoștiințele biologice din baze de date, precum Atlas of Genetics and Cytogenetics în Oncology and Haematology [31], care conține informații legate de toate adnotările de gene și Human Protein Atlas [21], au fost folosite pentru o mai bună înțelegere cu privire la structura, localizarea și caracteristicile genelor respective, toate fiind susceptibile a suferi mutații în cancerul de sân. În plus, informațiile clinice despre cancerul de sân din cadrul Catalogului de mutații somatice în cancer (COSMIC) [32] au fost utilizate pentru validarea genelor identificate și analizate de noi ca având legătură cu fenotipul de interes din modulul ales a fi modelat, iar diferite baze de date ca GEO, COSMIC, Uniprot sau Protein Atlas au fost folosite pentru a analiza caracteristicile fiecărei gene în parte și în ce mod acestea au legătură sau nu cu proliferarea celulelor alterate în organism.

În cadrul lucrării, a fost efectuată și o analiză de supraviețuire a genelor hub identificate în Capitolul 3 folosind plotterul Kaplan Meier. Toate genele analizate în rețeaua genică aleasă, au fost identificate în aplicația plotterul Kaplan Meier, ca biomarkeri biologici, unul din scopurile principale ale acestei aplicații fiind cel de validare a biomarkerilor identificați prin diverse tehnici.

Multe lucrări afirmă faptul că gena ADAM28 este supraexprimată în mai multe tipuri de cancere, printre care și cancerul de sân [33, 34], însă niciuna din analizele microarray tratate în lucrările disponibile din spațiul public nu a identificat gena ADAM28 ca posibilă genă target în cancerul de sân. În schimb, analiza noastră, pe lângă faptul că a selectat ca și cauză primară a modulului ales, un marker pronostic [35] (a cărei prezență și modificare a concentrației este corelată cu dezvoltarea de tumori) și anume gena RAC2, a identificat și gena efect cu rol de comunicator cu celulele imunitare ale corpului - ADAM28 și posibile căi de transmitere a mutațiilor între mai multe gene participante la procese de creștere și proliferare celulară din cadrul organismului uman. Astfel, obiectivul principal al tezei de doctorat, cel de realizare a unor

posibile trasee prin care eventualele mutații ce pot avea loc la nivelul genelor, să ducă la apariția cancerului de sân, a fost atins.

Accentul a fost pus cu precădere pe modelarea rolului genei efect ADAM28 în cadrul rețelei, în scopul concluzionării asupra posibilelor motive de apariţie şi promovare a cancerului de sân, fiind cunoscut faptul că mutații în ADAM28 induc progresia cancerului prin diferite mecanisme. Pe de altă parte, gena este considerată o potențială țintă terapeutică, fiind cunoscută, în stare normală, a juca un rol protector împotriva diseminării celulelor canceroase prin promovarea celulelor T [36, 37], astfel că o mai bună înțelegere a interacțiunilor din rețeaua din care face parte este vitală pentru stabilirea unor proceduri corecte de administrare a unor eventuale tratamente care să asigure buna funcționare a genei în organism.

Capitolul 5. Sinteza contribuțiilor științifice aduse și perspective de dezvoltare viitoare

5.1. Sinteza contribuțiilor științifice aduse

Cancerul de sân este una dintre principalele cauze de deces ale femeilor din lumea întreagă. Scopul lucrării a fost de a crea și modela rețele genice care au relevanță biologică în acest tip de cancer. Interpretarea valorilor expresiilor genice a constat într-o primă fază într-o analiză de tip microarray în limbajul R, urmată de o a doua faza, cea a analizei și validării genelor identificate și căutării informațiilor biologice despre ele în spațiul public, ca în final, rețeaua genică ce conține genele cele mai corelate din punct de vedere al distanței, să fie analizată și modelată în scopul găsirii și interpretării unor posibile căi de apariție a mutațiilor în genele specifice cancerului de sân (pe lângă cele cunoscute în mediul larg: ex BRCA1, BRCA2) – Figura 5.1.

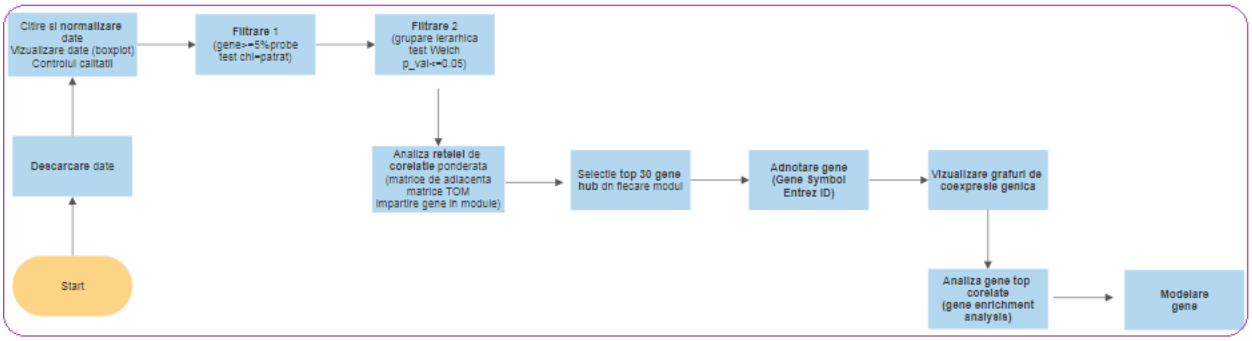

Figura 5.1. Pașii analizei microarray propuse

Pe scurt, pașii urmați au fost:

- ➔ instalarea pachetelor necesare analizei microarray în limbajul R: "BiocManager", "affy", "GEOquery", "affyPLM", "hgu133plus2.db", "AnnotationDbi.db", "annotate", "dynamicTreeCut", "WGCNA", "igraph";
- ➔ descărcarea fișierelor CEL din baza de date NCBI Gene Expression Omnibus (GSE48391 - matrici oligonucleotidice provenite de pe chipul Affymetrix Human Genome U133 Plus 2.0 Array);
- ➔ încărcarea și citirea datelor în limbajul de programare R;

Fișierele CEL de date brute ale fiecărui set de date microarray au fost importate în R (versiunea 4.1.0) folosind funcția ReadAffy a pachetului affy al BiocManager 3.13.

- ➔ normalizarea datelor cu metoda RMA, ce presupune trei pași: corectarea zgomotului de fond, normalizarea distribuțiilor și transformarea expresiilor în logaritmi în bază 2, doar pe baza valorilor PM din matrici;

Am ales o abordare diferită, legată de filtrarea numărului de gene, față de abordările uzuale. Abordarea mea este formată din două părți:

- ➔ filtrarea nr. 1 a setului de probe prin selectarea genelor exprimate în cel puțin 5% din probe, cu o variantă semnificativ diferită de cea mediană (test Chi-square);
- ➔ filtrarea nr. 2 a setului de probe: identificarea genelor exprimate diferențial pe baza valorilor p, utilizand un test statistic (Welch t-test);
- ➔ analiza rețelei de corelație ponderată, prin identificarea de module genice co-exprimate folosind pachetul WGCNA;
- ➔ adnotarea seturilor de gene la simboluri genetice (Entrez Gene ID și Symbol);

Adnotările și informațiile din spațiul biologic public au arătat că în toate din cele 13 module au fost găsite gene cu relevanță biologică în mai multe tipuri de cancer, iar în 5 (grey, greenyellow, yellow, purple, turquoise) din cele 13 module au fost identificați markeri care prezic apariția cancerului de sân.

- ➔ identificarea celor mai conectate gene din fiecare modul (gene cu conectivitate intramodul ridicată), adică acele gene hub cel mai puternic corelate cu procesul ales (în cazul nostru cancerul de sân);
- ➔ realizarea de grafuri de gene/ rețele de co-expresie genică folosind pachetul igraph (gene regulatory networks) pentru cele mai corelate 30 de gene din fiecare modul;
- ➔ identificarea caracteristicilor genelor cu rol în dezvoltarea cancerului de sân din toate modulele genice rezultate;
- ➔ analiza și caracterizarea unui singur modul genic pe baza grupării ierarhice (și de altfel cu cea mai mare relevanță biologică pentru studiul nostru), modul ce conține gene cu profiluri de expresie similare și funcții importante, cu relevanță biologică în progresia cancerului de sân;

Cu ajutorul mai multor baze de date publice am extras cunoștințe biologice despre gene și am verificat relevanța biologică a genelor din modulul ales, efectuând o analiză biologică (gene enrichment analysis) pe rețeaua obținută, pentru a vedea dacă modulul are sau nu sens biologic.
Informațiile găsite au fost pe scurt:

- o ARHGAP9: activatorul GTPase joacă roluri esențiale în reglarea creșterii celulare, diferențierea celulară, migrarea celulelor, având legătură cu Cdc42, o proteină implicată în reglarea ciclului celular și RAC1, care reglează mai multe căi de semnalizare ce controlează organizarea, transcripția și proliferarea celulară;
- o CSF2RB este receptor pt GM-CSF, un factor de creștere care induce diferențierea și proliferarea în măduva osoasă;
- o RAC2 are rol în a regla răspunsurile celulare, cum ar fi procesele celulelor apoptotice și epiteliale, fiind implicată în mutații;
- o CLEC12A codifică proteine cu rol în semnalizarea celulară și în răspunsul imun (unele dintre proteine fiind localizate în regiunea „genei ucigașe” de pe cromozomul 12p13);
- o STK17A este membru al apoptozei celulare;

- ADAM28 este cunoscută ca având legătură cu celulele sistemului imunitar, cu rol protector prin promovarea celulelor T în organism;

➔ vizualizarea proteinelor codificate de genele selectate și a unor regiuni specifice, precum localizarea mutațiilor ce pot apărea în interiorul acestora;

➔ modelarea modulului genic/ rețelei de co-expresie cu ajutorul inferenței probabilistice Bayes, sub forma unui graf aciclic direcționat (rețea bayesiană);

➔ identificarea unor posibile căi/trasee de transmitere a mutațiilor până la gena ADAM28, genă cu semnificație biologică de legătură cu celulele imunitare ale organismului;

➔ identificarea unor reguli de apariție a mutațiilor în gena ADAM28, în funcție de tipul de conexiune din rețeaua bayesiană și de tipul de informație furnizată (sau nu) asupra unei anumite gene din sistem;

Obiectivul principal al tezei de doctorat a fost de a identifica genele a căror expresie este corelată cu o anumită trăsătură fenotipică, în cazul nostru cancer de sân și care pot fi utilizate ca instrumente pentru realizarea unor posibile trasee prin care mutații ale genelor exprimate diferențiat să ducă la apariția acestei neoplazii. Astfel, rețelele bayesiene au fost alese pentru modelarea rețelei de co-expresie genică, acestea fiind utile în modelarea raționamentului în condiții de incertitudine. În plus, au un limbaj intuitiv și flexibil pentru reprezentarea dependențelor și independențelor între variabilele modulului ales. Ambele componente ale rețelei bayesiene, cantitativă și calitativă sunt transparente, în sensul deținerii complete a informațiilor asupra valorilor probabilităților și a observării continue a dependențelor între noduri, făcând schema genică foarte sugestivă. Informații valoroase pot fi extrase prin intermediul acesteia și al valorilor aplicate, în sensul analizei (inter)dependențelor între noduri. În afară de aceste caracteristici, rețelele bayesiene permit introducerea de mai multe tipuri de raționamente, față de alte tipuri de sisteme care nu permit decât unul singur.

În funcție de existența unui anumit tip de evidență sau informație (teste diagnostice, ecografii, prelevare sânge etc) asupra genelor, putem să oferim informații suplimentare legate de modificările acestora în cadrul procesului biologic ales. Astfel, dacă nu există informații certe de la medic sau expert, modelarea rețelei bayesiene se efectuează raportat la probabilitățile apriorice cunoscute anterior (de regulă calculate cu ajutorul formulei probabilității totale), rezultând noi probabilități marginale apriorice. Dacă aceste informații certe se cunosc, sunt calculate

probabilitățile aposteriorice ale nodurilor. În cazul nostru, a fost analizată influența informației certe asupra fiecărui nod din sistem, rezultând un număr de șapte cazuri de inferență probabilistică ale raționamentului cauzal sau predictiv:

a. Dacă informația legată de gena „părinte" ARHGAP9 este certă, atunci:

➔ probabilitățile celorlalte două gene „părinți" nu se modifică;

➔ probabilitatea genei CLEC12A se modifică;

➔ modificarea suferită de gena CLEC12A duce la modificări în genele STK17A și ADAM28;

➔ probabilitățile celorlalte două gene „părinți" nu se modifică;

b. Dacă informația legată de gena „părinte" CSF2RB este certă, atunci:

➔ probabilitățile celorlalte două gene „părinți" nu se modifică;

➔ probabilitatea genei CLEC12A se modifică;

➔ modificarea suferită de gena CLEC12A duce la modificări în genele STK17A și ADAM28;

c. Dacă informația legată de gena „părinte" RAC2 este certă, atunci:

➔ probabilitățile celorlalte două gene „părinți" nu se modifică;

➔ probabilitatea genei CLEC12A se modifică;

➔ modificarea suferită de gena CLEC12A duce la modificări în genele STK17A și ADAM28;

d. Dacă informațiile legate de două sau toate genele „părinți" RAC2, CSF2RB, ARHGAP9 sunt certe, atunci:

➔ probabilitatea genei CLEC12A se modifică și crește semnificativ față de cazurile a., b., c. (P(CLEC12A | RAC2=da, CSF2RB=da, ARHGAP9 =da));

➔ modificarea suferită de gena CLEC12A duce la modificări în genele STK17A (probabilitățile scad față de valorile inițiale) și ADAM28 (probabilitățile cresc față de valorile inițiale);

e. Dacă informația legată de gena CLEC12A este certă, atunci:

➔ probabilitățile genelor „părinți" cresc față de valorile inițiale;

➔ gena STK17A se modifică în funcție de (P(STK17A| CLEC12A=da));

➔ gena ADAM28 se modifică în funcție de (P(ADAM28| STK17A, CLEC12A=da));

- ➔ orice informație (certă) adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9 nu modifică încrederea în gena CLEC12A;
- ➔ orice informație adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9 nu modifică probabilitatea marginală a genei STK17A (genele inițiale și gena STK17A sunt independente condițional).
- ➔ orice informație adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9 nu modifică încrederea (probabilitatea marginală) în gena ADAM28, astfel că putem afirma faptul că genele inițiale și gena finală sunt independente condițional.

f. Dacă informația legată de gena STK17A este certă, atunci:

- ➔ probabilitățile genelor inițiale scad față de valorile inițiale;
- ➔ probabilitatea genei CLEC12A scade față de valoarea inițială P(CLEC12A);
- ➔ gena ADAM28 se modifică în funcție de (P(ADAM28| STK17A=da, CLEC12A));
- ➔ orice informație (certă) adusă asupra uneia sau mai multor gene „parinți” (RAC2, CSF2RB, ARHGAP9) modifică încrederea în gena CLEC12A, iar probabilitatea crește față de cea inițială;
- ➔ modificarea suferită de gena CLEC12A duce și la modificări în gena ADAM28 (creșterea probabilităților);

g. Dacă informațiile legate de ambele gene CLEC12A și STK17A sunt certe, atunci:

- ➔ gena ADAM28 se modifică în funcție de (P(ADAM28| STK17A=da, CLEC12A=da));
- ➔ orice informație (certă) adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9 nu modifică încrederea în genele CLEC12A și STK17A;
- ➔ orice informație (certă) adusă asupra genelor „parinți” RAC2, CSF2RB, ARHGAP9 nu modifică încrederea în gena ADAM28, astfel că putem afirmă faptul că genele inițiale și gena finală sunt independente condițional.

Construirea de modele intuitive grafic a devenit o tehnică populară care ne permite să înțelegem mai bine diferite procese, în cazul nostru o posibilă cale biologică implicată în rețelele de cancer de sân. Unul dintre cele mai importante avantaje ale acestor grafuri este versatilitatea modelelor, cu mențiunea respectării fiabilității sistemului.

Toate cazurile de mai sus aparțin raționamentului predictiv sau cauzal, în care presupunem cunoscute cauzele și vrem să vedem care este influența lor asupra genei efect (care e cauza X care a produs efectul Y). Noțiunea de cauză trebuie interpretată în sensul unui factor

care poate duce la creșterea sau descreșterea probabilității de realizare a altor parametrii pe care ii condiționează.

Abordarea descrisă în prezenta lucrare este utilă în descoperirea și înțelegerea cât mai completă a relațiilor cauzale dintre genele identificate ca find exprimate diferențial din setul de date GSE48391. A sesiza influența relațiilor de dependență și independență asupra unor ansambluri genice poate fi un factor definitoriu în prezicerea comportamentelor viitoare ale acestor gene sau pentru găsirea de relații noi între acestea. În această teză, toate genele participante în rețeaua bayesiană analizată au fost validate de baze de date biologice ca având o anumită legătură cu proliferarea celulară, celulele imunitare sau alte căi susceptibile a fi ținta unor alterări celulare periculoase asupra bunei funcționări a organismului uman. Modelul creat în mediul de dezvoltare Netica a respectat rețeaua de co-expresie genică oferită de pachetul igraph și aleasă de noi ca având cea mai mare relevanță biologică pentru studiu, păstrând un singur sens de legătură al genelor, începând cu genele RAC2, CSF2RB și ARHGAP9 și terminându-se cu gena ADAM28, cunoscută ca o protează ale cărei mutații induc progresia cancerului prin diferite mecanisme de proliferare a celulelor canceroase. Analiza în Netica a presupus multiple inferențe probabilistice asupra legăturilor dintre nodurile din rețea, cu scopul interpretării posibilelor căi de succedare a mutațiilor, pornind de la nivelul de bază (genele-cauză) și continuând până la nivelurile inferioare (gena efect implicată în creșteri tumorale și metastaze).

Astfel, am putut concluziona asupra următoarelor puncte:

1. Dacă nu avem informații (certe) asupra genelor RAC2, CSF2RB, ARHGAP9, trebuie luate în calcul toate legăturile din sistem (acest pas este cel mai complicat, pentru că implică fiecare nod și legătură din sistem).

➔ Dacă nu avem nici informații (certe) asupra genelor CLEC12A sau/și STK17A, trebuie luată în calcul legătura dintre genele inițiale ale rețelei și gena finală, putându-se interveni direct asupra genelor inițiale (asupra ratei de mutații sau a analizării modificărilor aminoacizilor în structura țesutului mamar neoplazic a acestor gene, pentru a încerca oprirea sau perpetuarea mutațiilor ulterioare din rețea sau măcar micșorarea ratei de mutații).

➔ Dacă avem informații certe asupra uneia sau amândurora dintre genele CLEC12A și STK17A, atunci legătura dintre genele inițiale ale rețelei și gena finală nu mai sunt de interes (fiind independente condițional), și putem interveni (biologic, printr-o exprimare

silențioasă și în limite normale a mutațiilor, care să nu mai permită invadarea celulelor sănătoase de către cele canceroase) doar asupra părinților lui ADAM28.

2. Dacă avem informații (certe) asupra genelor RAC2, CSF2RB, ARHGAP9, atunci trebuie analizate riguros legăturile dintre cele trei gene-cheie din rețea: CLEC12A, STK17A și ADAM28.

➔ Dacă nu avem informații (certe) asupra genelor CLEC12A sau/și STK17A, orice modificare suferită de una dintre genele părinți (CLEC12A și STK17A) duce la modificări pentru gena copil (ADAM28). În acest caz, trebuie analizate ambele gene CLEC12A și STK17A, în sensul identificării zonelor critice unde au loc mutațiile care promovează proliferarea spre noi gene.

➔ Dacă avem informații certe asupra uneia sau amândurora dintre genele CLEC12A și STK17A, atunci trebuie intervenit direct asupra genei ADAM28, pentru a încerca limitarea creșterii numărului de mutații în genă și a proliferării celulelor alterate în alte segmente importante de ADN implicate în creșteri tumorale, în cazul nostru, cancerul de sân.

Ca și limite ale analizei și identificării eventualelor trasee ale mutațiilor în genele susceptibile în controlul și reglarea progresiei ciclului celular și a apoptozei, putem specifica faptul că probabilitățile nodurilor au fost alese aleator, date reale despre aceste gene având un impact mult mai puternic și doveditor asupra celor determinate de noi, însă din căutările noastre, acestea nu au putut fi găsite pe platforme sau baze de date publice, pentru o analiză cât mai reală.

De altfel, nici disponibilitatea unor lucrări care să ateste legăturile dintre genele analizate în această teza nu a putut fi identificată, rezultatele noastre fiind validate prin identificarea acelor gene ale căror valori sunt semnificativ statistice, cu relevanță biologică în creșterea celulară, ale căror mutații duc la multiple transformări în diverse tipuri de cancer, însă cu precădere în cancerul de sân. Observațiile și informațiile din interiorul lucrării, confirmate de literatura de specialitate, sunt separate pentru fiecare genă în parte și vin în ajutorul unei mai bune înțelegeri a sistemului. O validare a relațiilor genelor din modulul ales a fi analizat ar fi de apreciat, însă acest lucru ar fi posibil doar prin lucrul complex într-un laborator de genetică, sub o atentă supraveghere și un efort generalizat al unei echipe mixte de doctori, ingineri, chimiști și biologi.

Originalitatea acestei lucrări constă într-o analiză integrată în limbajul R a expresiilor de gene din fișierul GSE48391, diferită față de celelalte existente. Analiza începe printr-o dublă

filtrare a setului de gene, cu ajutorul a două teste statistice: Chi-square și Welch t-test, la sfârșitul cărora a rezultat un număr redus de gene exprimate diferențial cu semnificație statistică. Pe baza acestor gene, cu ajutorul analizei de corelație WGCNA, au fost identificate treisprezece module genice, din care, pe baza legăturii de corelație ale top 30 de gene din fiecare modul, au fost create rețele de co-expresie genică cu ajutorul pachetului igraph. Din totalul de rețele rezultate, a fost selectat modulul de co-expresie genică cu cea mai mare relevanță biologică (din toate cele identificate de noi) în cancerul de sân. Multiple baze de date precum COSMIC, UniProt, canSAR sau Protein Atlas, au fost folosite pentru a analiza caracteristicile fiecărei gene în parte și a vedea în ce mod acestea au legătură cu proliferarea celulelor alterate în organism. Reproducerea schematică a modulului ales a fost efectuată cu ajutorul rețelelor bayesiene, în mediul de dezvoltare Netica. În scopul studierii și identificării principalelor relații ce stau la baza aparițiilor mutațiilor în gene, diverse probabilități au fost atribuite aleator pentru varabilele nodurilor, supunând modelul la influențe diverse, cu scopul de a determina caracteristicile pe care sistemul le poate avea, înainte de a dezvolta cancer. Accentul a fost pus cu precădere pe modelarea rolului genei efect ADAM28 în cadrul rețelei, în scopul concluzionării asupra posibilelor motive de apariţie și promovare a cancerului de sân, fiind cunoscut faptul că mutații în gena ADAM28 induc progresia cancerului prin diferite mecanisme.

5.2. Perspective viitoare

Un prim lucru pe care ne-am propus să-l testăm pe viitor este o altă caracteristică importantă a rețelelor bayesiene, și anume posibilitatea intervenției în procesul de cauzalitate, cu situații în care se poate mări graful cu una sau mai multe variabile de intervenție (noi cauze). Spre exemplu, în cazul rețelei bayesiene actuale, se poate introduce un nou nod “părinte” al nodului CLEC12A, sub forma unui nod de tip "tratament" și pe baza acestuia, să analizăm care este influența acestuia asupra nodului-efect ADAM28 și ale celorlalte noduri, anume în ce mod se vor modifica increderile asupra variabilelor rețelei prin introducerea noii variabile “tratament”, văzută ca un factor de ameliorare a efectelor pe care le pot produce mutațiile asupra genei-copil, rezultând astfel o reactualizare a genelor de la nivelurile inferioare genei nou apărute.

În altă ordine de idei, comparația mai multor gene specifice unor tipuri diverse de cancer ne poate oferi informații de mare interes despre genele care se găsesc în unul sau mai multe tipuri de neoplazii, dar și cele care diferă de la un tip de cancer la altul.

De altfel, o altă perspectivă viitoare ar fi și identificarea genele exprimate diferențial dintr-un material alterat față de unul genetic sănătos și comparația lor cu scopul de a identifica diferențele dintre cele două materiale genetice.

Bibliografie selectivă

[1] N.G. Bourbakis, "Bio-imaging and bio-informatics", IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 33(5):726–727, 2003.

[2] R. Fuchs, "From sequence to biology: the impact on bioinformatics", Bioinformatics, 18(4):505–506, 2002.

[3] G.B. Singh, "Fundamentals of Bioinformatics and Computational Biology. Methods and Exercises in Matlab", Modeling and Optimization in Science and Technologies, Springer, Vol 6, ISBN 978-3-319-11402-6, 2015.

[4] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, "Molecular Cell Biology", 4th edition, New York: W. H. Freeman, ISBN-10: 0-7167-3136-3, 2000.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, "Molecular Biology of the Cell", 4th edition, New York: Garland Science, ISBN-10: 0-8153-3218-1ISBN-10: 0-8153-4072-9, 2002.

[6] J.Ernst, "Computational Methods for Analyzing and Modeling Gene Regulation Dynamics", School of Computer Science, Machine Learning Department, CMU-ML-08-110, 2008.

[7] J.M. Keith, "Bioinformatics Vol I. Data, Sequence Analysis and Evolution", Methods in Molecular Biology, Humana Press, ISBN: 978-1-58829-707-5, 2008.

[8] I.P. Androulakis, E. Yang, R.R. Almon, "Analysis of Time-Series Gene Expression Data: Methods, Challenges and Opportunities", The Annual Review of Biomed.ical Engineering, 9:3.1-3.24, 2007.

[9] P.Baldi, S. Brunak, “Bioinformatics. The machine learning approach”, Second Edition, MIT Press, 2001.

[10] M.R. Speicher, S.E. Antonarakis, A.G. Motulsky, “Vogel and Motulsky’s Human Genetics. Problems and Approaches 4th Edition”, Spinger, 2010.

[11] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, “Hidden Markov models for cancer classification using gene expression profiles”, Information Sciences 316, 293-307, 2015.

[12] J. J. Pasternak, “An introduction to Human Molecular Genetics. Mechanisms of Inherited Diseases”, Second Edition, WILEY-Liss, 2005.

[13] The R Project for Statistical Computing. https://www.r-project.org/.

[14] Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/.

[15] K. Taskova, “Introduction to Microarray Analysis”, Computational Biology and Data Mining Group, Faculty of Biology, Gutenberg Universitat. 2016.
https://cbdm.uni-mainz.de/files/2016/02/GE_microarrays.pdf.

[16] Chi-square test for variance.
https://www.empirical-methods.hslu.ch/decisiontree/differences/variance/1-13chi-square-test-for-variance/.

[17] Institute for Digital Research & Education Statistical Consulting.
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/.

[18] P. Langfelder, S. Horvath, "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics, 9: 559, doi: 10.1186/1471-2105-9-559, 2008.

[19] J. Tang, D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong, G. Wu, "Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis", Front. Oncol., https://doi.org/10.3389/fonc.2018.00374, 2018.

[20] S. Horvath, J. Dong, "Geometric Interpretation of Gene Coexpression Network Analysis", PLOS Computational Biology, https://doi.org/10.1371/journal.pcbi.1000117, 2008.

[21] The Human Protein Atlas. https://www.proteinatlas.org/.

[22] Hierarchical Clustering function.
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust.

[23] Top Open Source Tools For Bayesian Networks. https://analyticsindiamag.com/top-8-open-source-tools-for-bayesian-networks/.

[24] Netica software. https://www.norsys.com/.

[25] I. Mihu, "Tehnici de decizie şi diagnoză", Editura Universitară, 2008.

[26] R. Mihu, M. Voinescu, O. Arsene, ”Tehnici de decizie şi diagnoză. Aplicaţii”, Editura Universitară, 2009.

[27] A. Pavlopoulou, D.A. Spandidos, I. Michalopoulos, "Human cancer databases (Review)", Oncol Rep, 33(1): 3–18, doi: 10.3892/or.2014.3579, 2015.

[28] I.-O. Lixandru-Petre, “Tehnici de asistare a deciziei şi diagnozei. Indrumar de laborator”, MatrixRom Bucuresti, ISBN: 978-606-25-0563-9, MatrixRom Bucuresti, 2020.

[29] Expression profiling by array. GSE36295.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36295.

[30] Expression profiling by array. GSE102907.
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102907

[31] Atlas of Genetics and Cytogenetics in Oncology and Haematology.
http://atlasgeneticsoncology.org/.

[32] Catalogue Of Somatic Mutations In Cancer. https://cancer.sanger.ac.uk/cosmic/.

[33] Y. Mitsui, S. Mochizuki, T. Kodama, M. Shimoda, T. Ohtsuka, T. Shiomi, M. Chijiiwa, T. Ikeda, M. Kitajima, Y. Okada, "ADAM28 is overexpressed in human breast carcinomas: implications for carcinoma cell proliferation through cleavage of insulin-like growth factor binding protein-3", Cancer Res, 66(20):9913-20, DOI: 10.1158/0008-5472.CAN-06-0377, 2006.

[34] C. Gérard, C. Hubeau, O. Carnet, M. Bellefroid, N.E. Sounni, S.Blacher, G. Bendavid, M. Moser, R. Fässler, A. Noel, D. Cataldo, and N. Rocks, "Microenvironment-derived ADAM28 prevents cancer dissemination", Oncotarget; 9(98): 37185–37199,
doi: 10.18632/oncotarget.26449, 2018.

[35] Markeri tumorali. https://www.cdt-babes.ro/articole/markeri_tumorali_generalitati.php.

[36] C. Gérard, C. Hubeau, O. Carnet, M. Bellefroid, N.E. Sounni, S. Blacher, G. Bendavid, M. Moser, R. Fässler, A. Noel et al., “Microenvironment-Derived ADAM28 prevents cancer dissemination”, Oncotarget, 9, 37185–37199, 2018.

[37] C. Hubeau, N. Rocks, D. Cataldo, “ADAM28: Another ambivalent protease in cancer”, Cancer Lett., 494, 18–26, 2020.

Lista publicațiilor

1. Articole publicate în reviste indexate Web of Science

- ➔ I.-O. Lixandru-Petre, C. Buiu, „An integrated breast cancer microarray analysis approach”, U.P.B. Sci. Bull., Series C, Vol. 84, Iss. 2, ISSN 2286-3540, 2022;
- ➔ I.-O. Lixandru-Petre, C. Buiu, „Modeling breast cancer gene expression using bayesian networks”, U.P.B. Sci. Bull., Series C, ISSN 2286-3540, 2022 - acceptată pentru publicare;

2. Articole publicate în manifestări științifice indexate Web of Science

- ➔ I.-O. Petre, C. Buiu, “An integrated gene expression analysis approach”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Iași, DOI: 10.1109/EHB.2015.7391442, WOS:000380397900095, 2015;
- ➔ I.-O. Petre, C. Buiu, “A colon cancer microarray analysis technique”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Sinaia, DOI: 10.1109/EHB.2017.7995412, WOS:000445457500067, 2017;
- ➔ I.-O. Petre, “Rb protein dynamic modeling”, Proceedings of the IEEE E-Health and Bioengineering Conference (EHB), Sinaia, DOI: 10.1109/EHB.2017.7995485, WOS:000445457500140, 2017;
- ➔ I.-O. Lixandru-Petre, “Modeling a Bayesian Network for a Diabetes Case Study", Proceedings of the International Conference on e-Health and Bioengineering (EHB), Iași, IEEE Xplore, DOI: 10.1109/EHB50910.2020.9280179, WOS:000646194100054, 2020;
- ➔ I.-O. Lixandru-Petre, "A Fuzzy System Approach for Diabetes Classification", Proceedings of the International Conference on e-Health and Bioengineering (EHB), Iași, Romania, IEEE Xplore, DOI: 10.1109/EHB50910.2020.9279882, WOS:000646194100008, 2020;

3. Articole publicate în manifestări științifice indexate BDI

➔ I.-O. Petre, C. Buiu, "Microarray Gene Expression Analysis using R", International Conference on Advancements of Medicine and Health Care through Tehnology (Meditech), Cluj-Napoca, IFMBE Proceedings book series (volume 59), DOI: 10.1007/978-3-319-52875-5_74, 2016;

4. Articole prezentate la conferințe internaționale

➔ I.-O. Petre, "Classifying different subtypes of malignant processes based on gene expression analysis", The Christie International Cancer Careers Conference, The Christie School of Oncology, Manchester, DOI:10.13140/RG.2.2.36691.94242, 2015;