(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 137833 A2**

(51) **Int.Cl.**
C12Q 1/68 (2006.01);
G16B 40/00 (2019.01);
G16H 50/30 (2018.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2022 00333**

(22) Data de depozit: **15/06/2022**

(41) Data publicării cererii:
29/12/2023 BOPI nr. **12/2023**

(71) Solicitant:
• **ARTIFICIAL INTELLIGENCE EXPERT S.R.L.**, *STR.ALEXANDRU VLAHUŢĂ, BLOC LAMA C, NR.45, CLUJ-NAPOCA, CJ, RO*

(72) Inventatori:
• **FLOARES ALEXANDRU**, *STR. ALEXANDRU VLAHUŢĂ, BL. LAMA C, AP. 45, CLUJ-NAPOCA, CJ, RO;*
• **ZETY ADRIAN**, *STR.AUREL VLAICU, BL.17, AP.11, CLUJ-NAPOCA, CJ, RO*

(54) # SISTEM INTELIGENT DE ANALIZĂ AUTOMATĂ A DATELOR MICRORNA MICROARRAY

(57) **Rezumat:**

Invenţia se referă la un sistem şi la un algoritm de analiză a datelor microRNA brute, obţinute prin tehnologiile Microarray, cu ajutorul unui flux de lucru ce combină etape bioinformatice şi de inteligenţă artificială pentru a obţine biomarkeri relevanţi pentru problema investigată şi modele predictive performante pentru soluţionarea acesteia, cu aplicare în domenii precum zootehnie, agricultură, criminalistică şi medicină. Algoritmul de analiză, conform invenţiei, cuprinde următoarele etape:

-preprocesarea datelor brute miRNA microarray şi controlul calităţii, etapă ce include şi normalizarea datelor,

-obţinerea profilului de expresie al microRNA,

-adnotarea miRNA care constă în ataşarea denumirii fiecărui miRNA la ID-urile din tabelul cu expresia fiecărei molecule, obţinându-se datele de intrare pentru un expert AI care aplică în mod automat algoritmi de inteligenţă artificială care efectuează:

-analiza exploratorie şi curăţarea datelor prin verificarea formei datelor, identificarea eventualelor date lipsă şi procentul acestora, eventualele anomalii prezente şi efectuarea unei analize statistice sumare precum: distribuţia datelor, valori minime, maxime, medii, etc., şi ulterior verifică corelaţiile dintre variabilele prezente, putându-se observa în acest fel potenţialul ca unele variabile să fie mai importante în analiza şi construcţia modelelor predictive,

-preprocesarea datelor înainte de antrenarea modelelor,

-dezvoltarea modelelor predictive şi

-optimizarea parametrilor algoritmilor de modelare.

Revendicări: 1
Figuri: 2

Fig. 2

RO 137833 A2

Descriere

Sistem Inteligent de Analiza Automată a Datelor microRNA Microarray

Invenția se referă la un sistem și un algoritm de analiza a datelor microRNA brute, obținute prin tehnologiile Microarray, cu ajutorul unui flux de lucru ce combina etape Bioinformatice si de Inteligenta Artificiala, pentru a obține biomarkeri relevanti pentru problema investigată și modele predictive performante, pentru a o solutiona.

Dat fiind rolul important al miRNA de reglare posttranslationala a activității genelor în întreaga lume vie a pluricelularelor, descoperirea biomarkerilor și dezvoltarea de modele predictive este importantă în domenii precum zootehnia, agricultura, criminalistica, si medicina.

MicroRNA (miRNA) sunt molecule mici de ARN necodificatoare de proteine care joacă un rol reglator important în traducerea genelor prin degradarea sau blocarea unor molecule de ARN mesager. Acestea influențează numeroase procese biologice majore, incluzand diferentierea, proliferare, apoptoza, inflamatia și metabolismul[1]. Datorită rolului proeminent pe care miRNA îl joacă in expresia genica si functionalitatea normala a organismelor, nu e surprinzător faptul că expresia lor aberantă poate duce la o multitudine de boli incluzand cancerul, bolile neurodegenerative, diabetul, conditiile cardiace, disfuncții ale rinichilor și ficatului, alterari ale sistemului imun. In plus, pe lângă contribuția la cauza a numeroase boli, microRNA pot fi utilizate și în terapiile țintite și în generarea de biomarkeri pentru diagnosticarea precoce a multor afectiuni. Dintre acestea, un rol important îl joacă moleculele de miRNA circulant care pot prelua rolul de biomarkeri[2] fiind detectabile din probe de sange, stand la baza metodelor neinvazive de detecție a numeroase boli [3].

Problema fundamentală intalnita in toate domeniile de aplicabilitate a analizei miRNA ce este departe de-a fi satisfăcător rezolvata, este aceeași. Utilizarea statisticii conventionale conduce la o lista de miRNA ce au o expresie diferită în două sau mai multe situații de interes distincte (expresie diferentiala), de exemplu Cancer sau Normal. O astfel de lista este de foarte mica utilitate practica. Inteligenta Artificiala (utilizata în sistemul nostru) face posibilă obținerea unor modele predictive, cu valoare practica (de exemplu teste de diagnostic al cancerului cu acuratețe >95%). Simpla aplicare a AI, chiar a celor mai avansate tehnici, nu conduce automat la astfel de performanțe intrucat sistemele vii au o proprietate aparte - redundantă funcțională. Acesta este de regula neglijata (nu și de către sistemul nostru) conducând la modele ce își

---

[1] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front. Endocrinol. 9, 402 (2018).

[2] Fabbri M (2010) miRNAs as molecular biomarkers of cancer. Expert Rev 10(4):435–444

[3] Chen X et al. (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 18(10):997–1006

degradează performanța cand sunt aplicate la cazuri noi, diferite de cele folosite la învățare. Exista numeroase domenii ce pot beneficia de pe urma analizei miRNA precum zootehnia, agricultura, medicina.

Asadar, printre principalele avantaje ale sistemului nostru se enumera și faptul ca iau în considerare redundantă funcțională conducând la modele robuste ce generalizează bine la noi cazuri.

În oricare din domeniile de aplicabilitate de mai sus, se doresc soluții pragmatice și cu o precizie cat mai inalta. Dezavantajul major al soluțiilor tehnice cunoscute, implementate in pachete software comerciale (precum Agilent miRNA Microarray Software, Partek Flow, sau open source (cum ar fi diverse pachete prezente în libraria R Bioconductor, de exemplu, AgiMicroRna, sau pachete individuale precum Chipster, este o consecinta a utilizarii statisticii conventionale, ce conduce doar la o lista de miRNA ce au o expresie cantitativă diferită în două sau mai multe situații de interes distincte (expresie diferentiala), de exemplu Cancer sau Normal. O astfel de soluție tehnica este de foarte mica utilitate practica intrucat:

- Metoda statistica de obținere a subsetului de miRNA relevanti pentru problema data, din întregul set de miRNA cunoscuți și determinati cantitativ de echipamentul microarray de laborator, nu primește feedback de la acuratețea cu care este rezolvata problema (de exemplu, diagnosticarea cancerului). Aceasta are impact asupra acuratetei soluției.
- Conține un număr foarte mare de miRNA exprimati diferit. Determinarea cantitativa a tuturor acestora, pentru aplicarea la situații concrete, este atat costisitoare cat și inutilă.
- Pentru reducerea costurilor se alege un prag arbitrar de expresie diferentiala pentru a scurta lista.
- Expresia biomarkerilor miRNA, obtinuti prin trunchierea listei complete trebuie apoi comparata, mai mult sau mai puțin manual, cu expresia acelorași miRNA la un nou caz la care se aplica.
- Drept rezultat, soluțiile tehnice cunoscute au adesea acuratețe scăzută, sunt costisitoare și dificil de aplicat.

Desi etapele de procesare bioinformatica a datelor miRNA microarray sunt similare la toate soluțiile (inclusiv a noastra), fiind standard, invenția noastra înlătura toate aceste dezavantaje ale solutiilor tehnice cunoscute, înlocuind analiza expresiei diferențiale (statistica conventionala) cu un întreg flux de lucru bazat pe Inteligenta Artificiala (AI). Mai mult, pentru a putea fi folosită cu usurinta și de către cei fără cunoștințe de programare sau AI, face apel la metode de automatizare a Machine Learning (AutoML). Datorita integrarii unui flux AI automatizat ce face inventia utilizabila si de catre utilizatorii fara cunostinte de AI sau programare am denumit pipeline-ul AI “Expert AI”.

Etapele din cadrul Expertului AI avand proprietarea de a se desfasura automat prin aplicarea AutoML. Toate acestea fac posibilă obținerea unor biomarkeri relevanti si a unor modele predictive cu valoare practica (de exemplu teste de diagnostic al cancerului cu acuratețe >95%).

Problema tehnică pe care o rezolvă invenţia este de a descoperi cat mai ușor, din datele brute miRNA microarray, biomarkeri și modele predictive cu o înaltă acuratețe, la costuri pe cat posibil mai reduse, ce pot soluționa probleme complexe din domeniile amintite (de exemplu, diagnosticarea cancerului). Aceste avantaje sunt posibile datorită înlocuirii metodei statistice de analiza a expresiei diferențiale cu un flux de lucru bazat pe Inteligenței Artificiala Automata (AutoML).

Pentru descoperirea biomarkerilor miRNA relevanti si transformarea acestora în instrumente pragmatice, de soluționare a problemelor în domeniile menționate mai sus, prin intermediul unor modele predictive, introducem metodologii de Inteligenta Artificiala adaptate specificului sistemelor vii, caracterizate prin redundanta functionala la nivel molecular (miRNA aici), în locul analizei statistice conventionale - expresia diferentiala a miRNA.

Se dă, în continuare, un exemplu de realizare a invenţiei, în legătură cu figurile 1, 2 şi 3, care reprezintă:

- Figura 1. Sistem - vedere de ansamblu
- Figura 2. Workflow Bioinformatic specific datelor Microarray si Expert AI

Se dă, în continuare un exemplu de realizare a invenției pe un set de date publice miRNA circulante microarray GSE168227 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE168227) a 48 pacienti cu 30 Carcinom celular scuamos oral si 18 Normal. Este importantă discriminarea cancerului de normal, motiv pentru care grupam cazurile de cancer în clasa Cancer (30 pacienti cu cancer) și clasa normal (18). Trebuie subliniat foarte clar ca pașii necesari obținerii datelor - colectarea sangelui/ plasmei, extractia ARN, pregatirea ADN complementar (cDNA), alegerea si utilizarea platformei de Microarray - NU fac obiectul invenției. Invenția noastra se aplica DOAR datelor nu si probelor pacientilor. Etapele realizarii invenției sunt următoarele:

Preprocesarea datelor si controlul calitatii. Acest pas include normalizarea datelor brute ceea ce verifica aparitia variatiilor tehnice intre array-urile din cadrul experimentului. Scopul normalizarii este de a indeparta variatia cauzata de motive tehnice pastrand variatia biologica a datelor. Procesul incepe cu vizualizarea datelor brute in vederea controlului calitatii, exista numeroase pachete software disponibile pentru realizarea acestui pas. Tot in cadrul acestui pas se vor identifica si inlatura coloanele cu valori nule ("NA"). O logaritmare in baza 2 a datelor poate fi realizata pentru a normaliza valorile de expresie pentru a urma o distributie Gaussiana.

Obtinerea profilului de expresie al microRNA. Pentru acest pas exista numeroase pachete comerciale si open-source disponibile, dupa cum au fost mentionate si anterior. Datorita naturii pachetelor open-source, acestea sunt mai des reinoite de catre comunitatile care le intretin fata de pachetele comerciale, astfel o strategie tipica pentru analiza bioinformatica a datelor obtinute cu tehonologii mature precum microarray este utilizarea unor pachete software open-source (ex: cele prezente in R Bioconductor) in construirea unui pipeline custornizat.

Adnotarea miRNA. Tabelul cu profilul de expresie al miRNA necesita legarea ID-urilor fiecarei molecule determinate de numele miRNA specific platformei si bazei de date utilizata de echipament (ex: miRBase). Astfel, e necesar un pas de atasare al denumirii fiecarui miRNA la ID-urile din tabelul cu expresia fiecarei molecule. Dupa acest pas se obtine input-ul ce va fi introdus in expertul AI.

Analiza Exploratorie și Curatarea Datelor. Primul pas al workflow-ului Expertului AI este analiza exploratorie a datelor. Aceasta începe cu verificarea formei datelor în care se observa numărul de linii și coloane. Una din aceste coloane va deveni ținta pentru care va fi realizata predicția utilizand algoritmii AI. Datele de input și output trebuie sa aiba formatul corect - coloanele miRNA conțin numere reale, iar ținta - diagnosticul aici - este variabila categoriala.

Se vor identifica eventualele date lipsa și procentul acestora, eventualele anomalii prezente în date și se va proceda la o analiza statistica sumara (distributia datelor, valori minime, maxime, medii, etc.). Ulterior se vor verifica corelatiile dintre variabilele prezente, astfel se poate observa potențialul ca unele variabile sa fie mai importante în analiza și construcția modelelor predictive.

Preprocesare datelor inaintea de antrenarea modelelor:

Identificare valorilor lipsa și imputarea acestora cu metode avansate.

Identificarea de Outliers prin multiple metode și îndepărtarea acestora.

Identificare de eventuale denumiri multiple pentru aceeași coloana și corectare.

Logaritmarea în baza 2 a datelor. Se considera ca distribuția datelor se apropie de cea Gausiana, prin aceasta transformare.

Normalizarea. Se alege una din următoarele metode: scalarea datelor între 0 și 1, sau intre –1 si 1 sau, mai frecvent, standardizarea – media datelor devine 0 iar deviatia standard devine 1.

Dezvoltarea modelelor predictive

Acest pas începe cu inițializarea interfeței Expertului AI care va rula în cadrul unei instanțe de tip notebook (ex: Jupyter Notebooks). Prin utilizarea metodelor AutoML, modelele predictive vor fi dezvoltate automat. Totusi, vor fi cerute anumite setari din partea utilizatorului.

Se vor incarca librăriile și pachetele necesare mediului de lucru pentru Expertul AI.

Se vor defini perechile input - valorile de expresie a miRNA si output - coloana tinta (diagnostic) cu valorile "Cancer" si "Non-Cancer").

Se definesc seturile de antrenare si testare. Daca numărul cazurilor permite, se defineste si cel de al 3 lea set (de validare). Cel mai adesea, numărul cazurilor este relativ mic, motiv pentru care se definesc doar un set de antrenare și testare și se utilizează cross validation pentru antrenare. De exemplu, se împart cazurile de antrenare în 10 subseturi, învățarea are loc pe 9 dintre ele iar testarea pe al 10 lea, substerile fiind permutate.

In acest exemplu, performanta clasificarii binare Cancer - Non-Cancer poate fi reprezentată de mai multor metrici precum Acuratețea, AUC (Area Under the Curve), Precizia, Kappa si MCC (coeficient de corelatie Matthews).

Se antrenează toate modele disponibile in librariile incarcate si se evaluează performanța acestora, rezultand o ierarhie de modele.

Optimizarea parametrilor algoritmilor de modelare:

Orice algoritm de modelare are el însuși o serie de parametri (de exemplu, adancimea arborilor decizionali). Valorile obișnuite (default) ale acestora pot sa nu conducă la performanta maxima posibila, dar pot fi optimizate. În acest scop se utilizează subsetul de validare care, cand sunt putine cazuri, poate fi un subset al setului de antrenare. Daca performanța obținută este superioară se retin valorile corespunzătoare ale parametrilor algoritmului

Prin aplicarea invenției se obțin următoarele avantaje:

- Se descoperă biomarkeri cu adevărat relevanți, intrucat acestia sunt doar o componenta a unor modelelor predictive a caror performanta este masurata. Daca modelul bazat pe un set de biomarkeri are o acuratete de peste 95%, in mod clar markerii sunt relevanti. Daca, insa, acuratețea este doar de 60% de exemplu, valoare biomarkerilor descoperiți este cvasi nula, intrucat alegerea clasei la întâmplare are o acuratețe de 50%, destul de apropiata.
- Se descoperă modele predictive de mare performanta, intrucat acestea extrag toata informația conținută în inputuri, relativ la outputuri, fiind și direct aplicabile la noi cazuri, spre deosebire de simplele liste.
- Îmbunătățiri semnificative ale eficientei însoțite de scăderea costurilor în domeniile de aplicabilitate.

Revendicari

Inventia permite descoperirea de biomarkeri și modele predictive cu o înaltă acuratețe. Tehnicile AI din inventie elimina variabilele ce nu diferă cantitativ intre situațiile de discriminat pastrand doar variabilele informative pentru predicție, selectarea acestora se desfasoara utilizand un mecanism de feedback de la acuratețea predicției. In plus prin utilizarea abordarilor de tip AutoML se testeaza in mod automat numerosi algoritmi AI fapt care permite experimentarea și alegerea celui mai bun algoritm fata de abordarile de analiza statistica conventionala.

Desene

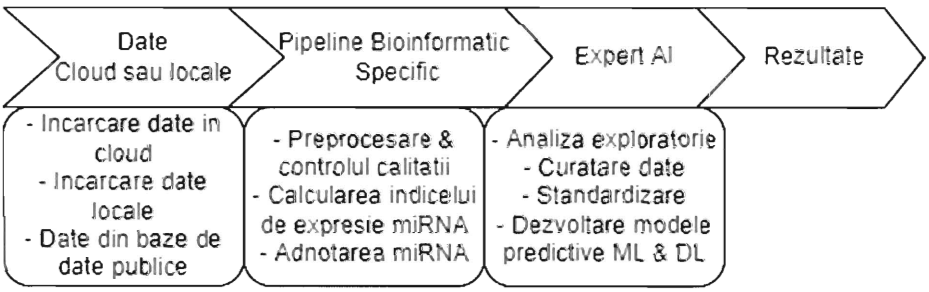

Figura 1. Sistem - vedere de ansamblu

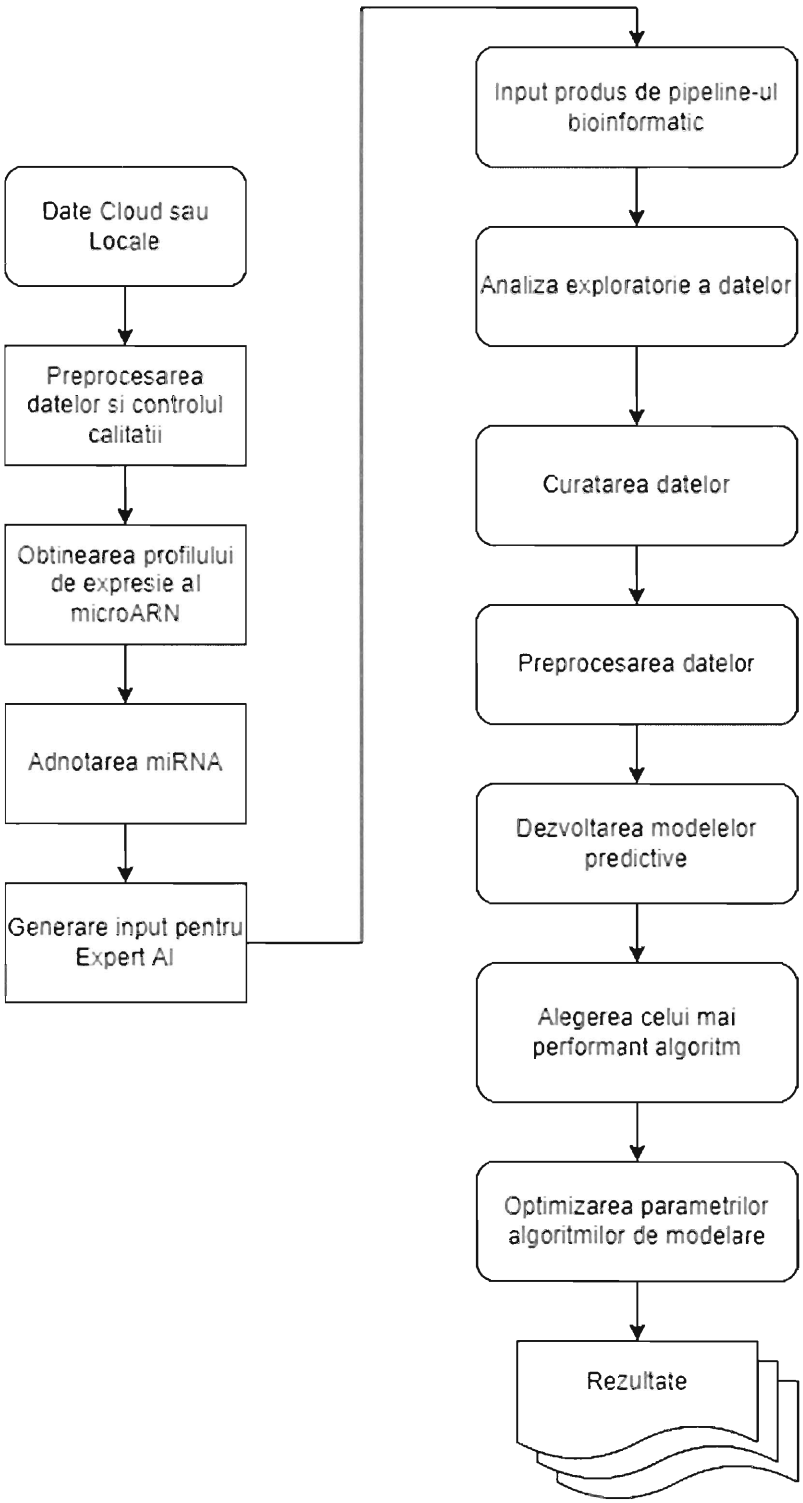

Figura 2. Workflow Bioinformatic specific datelor Microarray si Expert AI