ROMÂNIA

(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

(11) **RO 137834 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 00334**

(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 NGS

(57) **Rezumat:**

Invenţia se referă la un sistem şi la un algoritm de analiză a datelor microRNA brute, obţinute prin tehnologii NGS (Next Generation Sequencing), cu ajutorul unui flux de lucru ce combină etape bioinformatice şi de inteligenţă artificială pentru a obţine biomarkeri relevanţi pentru o problemă investigată şi modele predictive pentru a o soluţiona, cu aplicare în domenii precum zootehnia, agricultura, criminalistica şi medicina. Algoritmul de analiză, conform invenţiei, cuprinde următoarele etape:

-preprocesare şi controlul calităţii datelor, în care se verifică calitatea secvenţelor şi se produce un raport detaliat cu toate informaţiile din conţinutul unei probe analizate, iar raportul este analizat şi se decide asupra filtrării secvenţelor de calitate nesatisfăcătoare,

-îndepărtarea secvenţelor adaptor,

-alinierea secvenţelor pe o referinţă,

-cuantificarea miRNA, în care un algoritm va colapsa toate secvenţele identice şi le va număra, după care le va introduce într-o matrice care va fi completată ulterior cu numărul secvenţelor din fiecare probă, obţinând un profil complet de expresie al miRNA care este utilizat ca intrare într-un Expert AI al cărui flux de lucru cuprinde: analiza exploratorie şi curăţarea datelor, preprocesarea datelor înainte de antrenarea modelelor, dezvoltarea de modele predictive şi optimizarea parametrilor algoritmilor de modelare.

Revendicări: 1
Figuri: 2

Fig. 2

RO 137834 A2

RO 137834 A2

Descrierea

Sistem Inteligent de Analiza Automata a Datelor microRNA NGS

Invenția se referă la un sistem şi un algoritm de analiza a datelor microRNA brute, obtinute prin tehnologiile Next Generation Sequencing (NGS), cu ajutorul unui flux de lucru ce combina etape Bioinformatice si de Inteligenta Artificiala (AI), pentru a obține biomarkeri relevanti pentru problema investigată şi 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 si dezvoltarea de modele predictive este importanta in domenii precum zootehnia, agricultura, criminalistica, si medicina.

MicroRNA (miRNA) sunt molecule mici de ARN necodificatoare de proteine care joacă un rol reglator important în translatia genelor prin degradarea sau blocarea unor molecule de ARN mesager. Acestea influențează numeroase procese biologice majore, incluzand diferentierea, proliferare, apoptoza, inflamatia şi 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 şi ficatului, alterari ale sistemului imun. In plus, pe lângă contribuția la cauza a numeroase boli, microRNA pot fi utilizate şi în terapiile țintite şi 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].

În oricare din domeniile de aplicabilitate, problema fundamentală, ce este departe de-a fi satisfăcător rezolvata, este aceeaşi. 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,

---

[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

chiar a celor mai avansate tehnici, nu conduce automat la astfel de performanţe intrucat sistemele vii au o proprietate aparte - redundantă funcţională. Acesta este de regula neglijata (nu şi de către sistemul nostru) conducând la modele ce îşi degradează performanţa cand sunt aplicate la cazuri noi, diferite de cele folosite la învăţare. Exista numeroase domenii ce pot beneficia de pe urma analizei miRNA precum zootehnia, agricultura, medicina.

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

Datele brute miRNA nu vorbesc de la sine. Pentru a deveni instrumente practice utile, de exemplu un test de diagnostic al cancerului, etc., este necesara descoperirea din date a unor modele predictive cat mai performante (de exemplu, cu acuratete >95%). Acest lucru este posibil doar prin aplicare unor tehnici avansate de Inteligenta Artificiala (AI). Totusi, sistemele vii (plante, animale, fiinte umane) prezinta o caracteristica aparte ce face ca simpla aplicare a AI sa nu dea rezultate satisfăcătoare - modele predictive robuste care generalizează bine (acurateţea nu se degradează) cand sunt aplicate la cazuri noi, diferite de cele utilizate la învăţare. Sistemele vii au o înaltă redundantă funcţională. Simplu spus, aceeaşi funcţie se poate realiza la nivel molecular in mai multe moduri. Aceasta ofera sistemelor vii rezistenţă la o multitudine de alterari posibile. Ce mai adesea, redundanta funcţională este neglijată şi se încearcă obţinerea unui subset de miRNA cat mai mic.

In prezent, sistemele, algoritmii şi fluxurile de lucru pentru descoperirea de biomarkeri miRNA, implementate in pachete software comerciale - Qiagen CLC Workbench, Illumina miRNAs Analysis, Partek Flow, - sau open source - Chipster, COMPSRA, Prost!, sau DIANA-mAP - se finalizează prin analiza expresiei diferenţiale şi nu prin modele predictive performante. Se obţine doar o lista de miRNA exprimate diferit între cele două sau mai multe situaţii de interes (de exemplu, Cancer si Normal). Aceasta lista este apoi trunchiată arbitrar şi aplicată cazurilor noi. Se compara expresia miRNA a cazurilor noi cu aceasta lista pentru încadrarea în una din clasele investigate. Prin contrast, solutia noastra descoperă modele predictive. Expresia miRNA a noilor cazuri este introdusă în acestea şi se obţine incadrare in una din clase. Desi reteta de analiza ce proneste de la datele miRNA NGS brute este similara la toate soluţiile existente, fiind oarecum standard, solutia noastra introduce un întreg flux de analiza cu IA ce conduce la soluţii de înalt pragmatism şi performanţa.

Toate sistemele conventionale existente in prezent prezinta urmatoarele probleme:

- Metoda statistica de obținere a subsetului de miRNA relevanti pentru problema data, din întregul set de miRNA cunoscuți și determinati cantitativ de echipamentul NGS si de pipeline-ul bioinformatic, 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 aceloraşi miRNA la un nou caz la care se aplica.
- Drept rezultat, soluțiile tehnice cunoscute au adesea acuratețe scăzută, sunt costisitoare şi dificil de aplicat.

Desi etapele de procesare bioinformatica a datelor miRNA NGS sunt similare la toate soluțiile (inclusiv a noastra), fiind standard, invenția noastra înlătura dezavantajele 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 şi de către cei fără cunoştințe de programare sau AI, face apel la metode de automatizare a Machine Learning (AutoML). Am numit acest flux automat de aplicare a Inteligenței Artificiale Expert AI pentru ai sugera functionalitatea.

Toate acestea fac posibilă obținerea unor biomarkeri relevanti si a unor modele predictive cu valoare practica (de exemplu teste de diagnostic al cancerului cu acurateţe >95%).

În toate domeniile de aplicabilitate, problema importantă dar dificila este de a descoperi biomarkeri şi modele predictive cu o înaltă acuratețe, simpla analiza a expresiei diferențiale fiind total nesatisfăcătoare. De fapt, tehnicile de AI din sistemul nostru vor elimina oricum variabilele ce nu diferă cantitativ intre situațiile de discriminat intrucat acestea nu sunt informative pentru predicție. Aplicand AI, selectarea variabilelor se face cu ajutorul unui feedback de la acuratețea predicției. In plus exista o paleta larga de algoritmi AI ceea ce permite experimentarea şi alegerea celui mai bun spre deosebire de analiza statistica conventionala. 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. In plus, prin implementarea Inteligentei Artificiale Automate, inventia poate fi utilizata cu usurinta si de catre persoane fara cunostinte avansate de programare.

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 NGS si Expert AI

Se dă, în continuare un exemplu de realizare a invenției pe un set de date publice miRNA-Seq NGS de cancer la san preluate din The Cancer Genome Atlas (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) cu o cohorta de 1079 de pacienti cu cancer la san si un numar egal de cazuri non-cancer. Trebuie subliniat foarte clar ca paşii necesari obţinerii datelor - colectarea sangelui/ plasmei, extractia ARN, conversia ARN în librării miRNA pentru NGS şi secventarea probelor - NU fac obiectul invenției. Invenția noastra se aplica DOAR datelor nu si probelor pacientilor. Etapele realizarii invenției sunt următoarele:

Preprocesare şi controlul calității datelor. Se verifica calitatea secvențelor şi se produce un raport detaliat cu toate informațiile despre conținutul probei. Rapoartele pentru probe individuale sunt ulterior incluse într-un raport general ce conţine informaţii despre întregul experiment prezentand detalii precum conținutul de baze azotate (Adenina, Guanina, Citozina, Uracil), prezenta şi structura secvențelor adaptor (în situația în care acestea nu au fost îndepărtate anterior, proces ce este oferit de către unele centre de secventiere). Raportul este analizat şi se decide asupra filtrarii secvențelor de calitate nesatisfăcătoare.

Îndepărtarea secvențelor adaptor. Acestea pot fi cunoscute anterior, facilitățile de secventiere oferind de obicei o documentație în care menționează secvențele adaptor utilizate, sau pot fi identificate prin utilizarea unor algoritmi specializați. Deodata cu eliminarea secvențelor adaptor se poate introduce şi un prag de lungime pentru secvențele ce vor fi obținute. Dat fiind ca secvențele

miRNA au o lungime medie de 22 de nucleotide, se va seta un prag cu o valoare între 22-30 nucleotide pentru a păstra doar moleculele de interes şi pentru a scurta timpul de procesare in etapa de aliniere (se va utiliza o valoare mai mare de 22 pentru a acomoda eventualele erori rămase din etapa de filtrare).

Aliniere a secvenţelor pe o referinta. Se va utiliza cea mai recentă varianta publicata şi adnotată a genomului uman. Secvenţele preprocesate vor fi aliniate pe genom utilizând cei mai performanţi algoritmi de aliniere după care vor fi adnotate utilizand cea mai recentă versiune a bazelor de date pentru miRNA (ex: miRBase).

Cuantificarea miRNA. Un algoritm va colapsa toate secvenţele identice şi le va număra după care le va introduce intr-o matrice (tabel). Aceasta matrice va fi completata ulterior cu numărul secvenţelor din fiecare probă obtinand astfel un tabel cu profilul complet de expresie al miRNA. Profilul de expresie al miRNA e utilizat ca input in Expertul AI.

Analiza Exploratorie şi 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 şi coloane. Una din aceste coloane va deveni ţinta pentru care va fi realizata predicţia utilizand algoritmii AI. Datele de input şi output trebuie sa aiba formatul corect - coloanele miRNA conţin numere reale, iar ţinta - diagnosticul aici - este variabila categoriala.

Se vor identifica eventualele date lipsa şi procentul acestora, eventualele anomalii prezente în date şi 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 potenţialul ca unele variabile sa fie mai importante în analiza şi construcţia modelelor predictive.

Preprocesare datelor inaintea de antrenarea modelelor:

Identificare valorilor lipsa şi imputarea acestora cu metode avansate.

Identificarea de Outliers prin multiple metode şi îndepărtarea acestora.

Identificare de eventuale denumiri multiple pentru aceeaşi coloana şi corectare.

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

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

Dezvoltarea modelelor predictive

Prin utilizarea metodelor AutoML, modelele predictive vor fi dezvoltate automat. Totusi, vor fi cerute anumite setari din partea utilizatorului, facilitate de folosirea unui notebook (de exemplu, Jupyter Notebooks).

Se vor incarca librăriile şi pachetele necesare mediului de lucru pentru Expertul AI (fluxul de lucru 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 şi testare şi se utilizează cross validation pentru antrenare. De exemplu, se împart cazurile de antrenare în 10 subseturi, învăţarea 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 Acurateţea, AUC (Area Under the Curve), Precizia, Kappa si MCC (coeficient de corelatie Matthews).

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

Optimizarea parametrilor algoritmilor de modelare:

Orice algoritm de modelare are el însuşi o serie de parametri (de exemplu, adancimea arborilor decizionali pentru ansamblele de arbori decizionali). Valorile obişnuite (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 performanţa obţinută este superioară se retin valorile corespunzătoare ale parametrilor algoritmului.

Prin aplicarea invenţiei se obţin următoarele avantaje:

- Se descoperă biomarkeri cu adevărat relevanţi, 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, acurateţea este doar de 60% de exemplu, valoare biomarkerilor descoperiţi este cvasi nula, intrucat alegerea clasei la întâmplare are o acurateţe 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 şi modele predictive cu o înaltă acurateţe. Tehnicile AI din inventie elimina variabilele ce nu diferă cantitativ intre situaţiile de discriminat pastrand doar variabilele informative pentru predicţie, selectarea acestora se desfasoara utilizand un mecanism de feedback de la acurateţea predicţiei. In plus prin utilizarea abordarilor de tip AutoML se testeaza in mod automat numerosi algoritmi AI fapt care permite experimentarea şi alegerea celui mai bun algoritm fata de abordarile de analiza statistica conventionala.

Invenţia nostra consta într-un "Sistem Inteligent de Analiza a Datelor miRNA NGS" capabil de a prelua datele brute miRNA produse de echipamentele de laborator NGS, de a le preprocesa din punct de vedere bioinformatic (in modul standard) şi de a descoperi biomarkeri în scopul soluţionării diverselor probleme din domeniile de aplicabilitate (de exemplu, diagnosticul cancerului), caracterizată prin aceea că:

1. Metodelor statistice de analiza a expresiei diferenţiale sunt înlocuite cu un flux de lucru bazat pe Inteligenta Artificiala iar
2. Acest flux de lucru (Expert AI) este automatizat,

ceea ce facilitează dramatic descoperirea biomarkerilor relevanţi şi a unor modele predictive cu înaltă acurateţe, oferind soluţii pragmatice problemelor complexe din domeniile de aplicabilitate.

Desene

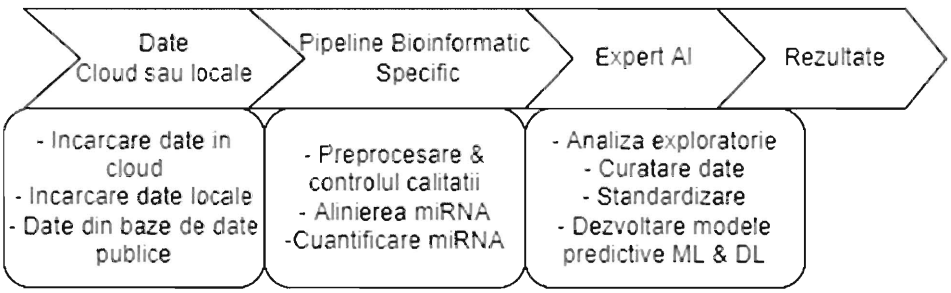

Figura 1. Sistem - vedere de ansamblu

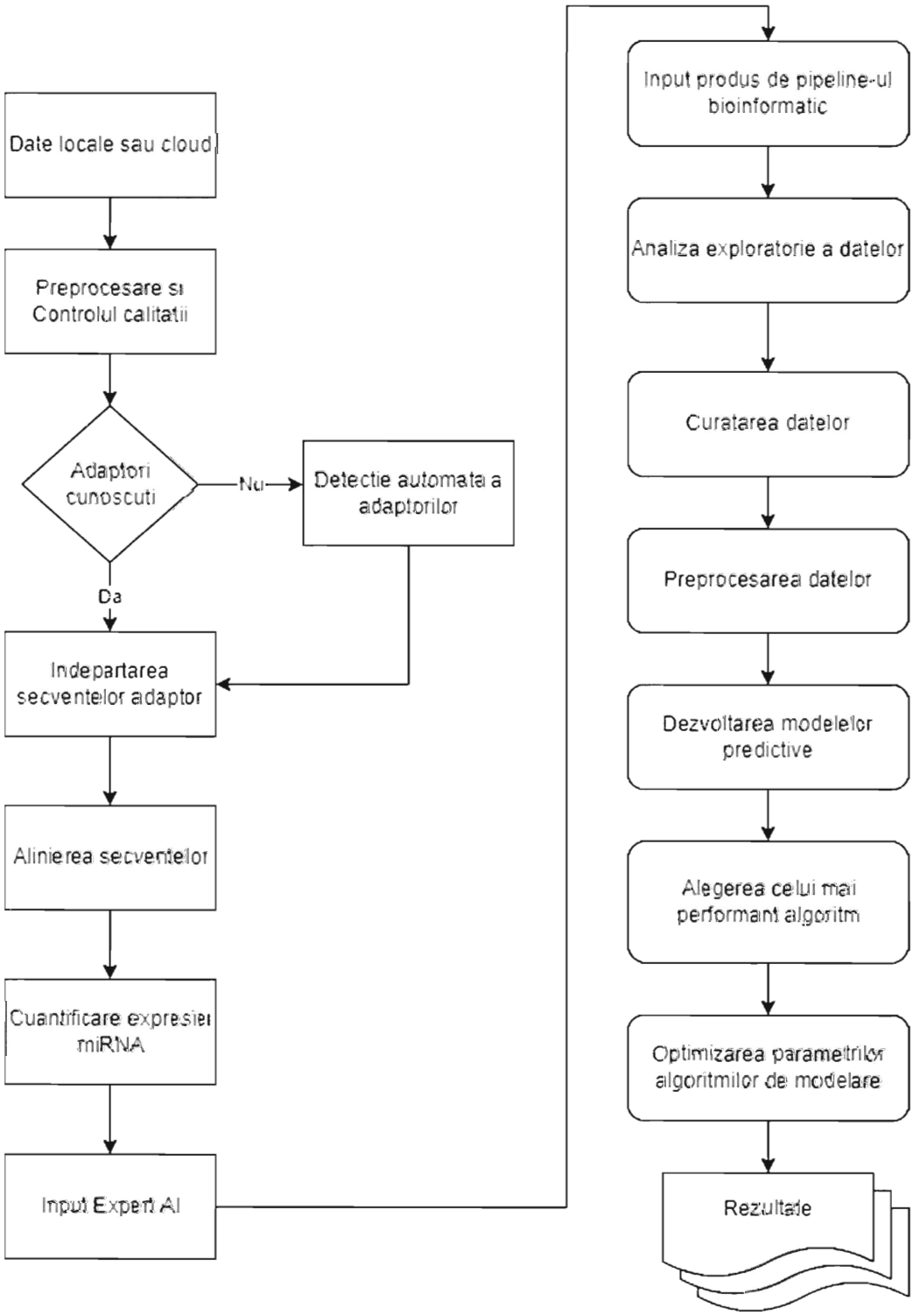

Figura 2. Workflow Bioinformatic specific datelor NGS si Expert AI