



UNIVERSITATEA TEHNICĂ A MOLDOVEI
ȘCOALA DOCTORALĂ

Cu titlu de manuscris
CZU: 004.9:519.21:575

MUNTEANU VIOREL

MODELE MARKOVIENE ÎN TIMP CONTINUU BAZATE PE
FILOGENIE PENTRU DINAMICA GENELOR ÎN PANGENOMURILE
MICROBIENE

122.03 Modelare, metode matematice, produse program

Rezumatul științific al tezei de doctor în informatică

Conducător științific:

Bostan VIOREL, doctor habilitat, profesor universitar

CHIȘINĂU, 2026

Teza a fost elaborată în cadrul Departamentului „**Inginerie Software și Automatică**”,
Facultatea Calculatoare, Informatică și Microelectronică a **Universității Tehnice a Moldovei**.
Școala doctorală a Universității Tehnice a Moldovei

Autor: MUNTEANU Viorel

Conducător Științific:

BOSTAN Viorel membru corespondent al AȘM, doctor habilitat în științe tehnice,
profesor universitar, Universitatea Tehnică a Moldovei, Republica

Comisia de susținere publică:

Președinte: GUȚULEAC Emilian, doctor habilitat în științe tehnice, profesor universitar,
Universitatea Tehnică a Moldovei, Republica Moldova

Membru: BOSTAN Viorel, membru corespondent al AȘM, doctor habilitat în științe tehnice,
profesor universitar, Universitatea Tehnică a Moldovei, Republica Moldova

Referent: MANGUL Serghei, doctor în informatică, Sage Bionetworks, SUA

Referent: LOBIUC Andrei, doctor în biologie, profesor universitar, Universitatea “Ștefan cel
Mare” din Suceava, România

Referent: CEPOI Liliana, membru corespondent al AȘM, doctor habilitat în biologie, profesor
universitar, Universitatea Tehnică a Moldovei, Republica Moldova

Susținerea tezei va avea loc la **30 Martie 2026, orele 09:00** în ședința Comisiei de Doctorat
din cadrul Școlii doctorale a **Universității Tehnice a Moldovei** (aprobată prin decizia
Consiliului Științific al UTM din 27 ianuarie 2026, proces-verbal nr. 1), str. Studenților 9/7,
Blocul de studii nr. 3 a UTM, Aula 3-3 „Amdaris”, Chișinău, MD – 2068, Republica
Moldova.

Teza de doctor și rezumatul științific pot fi consultate la Biblioteca Tehnico-Științifică a
Universității Tehnice a Moldovei și pe pagina web a ANACEC (www.anacec.md).

Autor:

VIOREL Munteanu 

Conducător științific:

BOSTAN Viorel, doctor habilitat în științe tehnice, profesor universitar



© Viorel Munteanu, Universitatea Tehnică a Moldovei, 2026

REPERELE CONCEPTUALE ALE CERCETĂRII

Actualitatea temei abordate. Dinamica evolutivă a conținutului genomurilor microbiene, determinată de procesul de acumulare de mutații, selecție, recombinare, transfer orizontal de gene și alte evenimente de duplicare și pierdere a genelor, stă la baza versatilității ecologice și a potențialului de adaptare al populațiilor microbiene [1–4]. Aceste procese generează fluiditate genomică extinsă, intensifică diversificarea funcțională și permit răspunsuri rapide la presiunile de mediu în ecosisteme naturale, cele antropice și cele asociate gazdei. Deși pangenomul a apărut drept un cadru conceptual puternic pentru captarea acestei variabilități genomice, abordările existente sunt în mare parte centrate pe culturi de izolate bacteriene, ceea ce le limitează aplicabilitatea în cazul speciilor de microbi rari sau necultivabili [5–7]. În consecință, atestăm o necesitate tot mai pronunțată de metode capabile să reconstruiască și să interpreteze pangenomul microbial direct din date de secvențiere metagenomice [8]. Satisfacerea acestei nevoi devine din ce în ce mai urgentă, având în vedere că microbiomul din mediu și cel asociat gazdelor sunt recunoscute drept rezervoare majore de rezistență antimicrobiană, inovație metabolică și potențial patogen [9]. Astfel, devine esențială dezvoltarea de metode bioinformatic de cuantificarea *in situ* a dinamicii conținutului de gene și implementarea sistemelor de supraveghere genomică urbană pentru monitorizarea genelor de rezistență antimicrobiană și determinanților de virulență [6, 9]. În ecosisteme antropice, inclusiv apele uzate, solurile și aerul urban, estimarea ratelor și mecanismelor fluxului genic (transferul orizontal de gene, recombinările, câștigul și pierderea de gene) devin critice pentru trasarea apariției trăsăturilor adaptive și pentru anticiparea răspunsurilor microbiene la presiunile selective [6, 10, 11]. Modelele de genomică comparativă au devenit esențiale pentru elucidarea acestor procese, oferind un cadru matematic riguros pentru modelarea evoluției trăsăturilor discrete prin procese Markov în timp continuu [12, 13]. Aceste modele permit identificarea evenimentelor evolutive de interes pe ramurile filogenetice, facilitând caracterizarea mecanismelor care determină dinamica evolutivă a caracterelor în cadrul arborelui filogenetic. Extinderea și adaptarea acestor metode comparative la seturi de date genomice și pangenomice devin tot mai necesare, întrucât permit surprinderea întregului spectru al diversității genomice, inclusiv a liniilor microbiene necultivabile. Prin integrarea contextului filogenetic cu variația genomică structurală și funcțională, aceste abordări facilitează investigarea forțelor evolutive care determină trăsăturile și funcțiile adaptive, precum și reorganizarea rapidă a repertoriilor genomice în comunitățile microbiene.

Importanța problemei abordate. Studiul evoluției genomurilor microbiene se situează la intersecția bioinformaticii, modelării matematice, analizei datelor de secvențiere și genomicii comparative [5–7]. Progresele rapide în tehnologiile de secvențiere au extins considerabil aria

pangenomicii computaționale, o disciplină emergentă axată pe analiza comparativă a conținutului genetic în populațiile microbiene de interes, inclusiv în cele care populează ecosistemele urbane [14, 15]. Integrarea metodelor comparative filogenetice cu analiza metagenomică permite identificarea tiparelor specifice de câștig și pierdere a genelor la nivel de linie evolutivă în sisteme microbiene insuficient reprezentate de genomuri de referință sau alcătuite din microorganisme care nu pot fi cultivate. Astfel de abordări integrative contribuie la o înțelegere mai amplă a dinamicii genomice în comunitățile naturale și susțin aplicațiile de supraveghere a sănătății publice, inclusiv a rezistenței anti-microbiene (RAM) sau a factorilor de virulență. Microbiomul urban reprezintă un context deosebit de provocator, din cauza plasticității genomice accentuate care influențează adaptarea microbiană, transmiterea și diseminarea RAM în mediile antropogenice dens populate [4, 9, 14, 15]. Analiza acestor procese într-un cadru evolutiv contribuie la înțelegerea modului în care funcțiile și trăsăturile microbiene emergente apar, persistă și se răspândesc în mediul urban. În pofida progreselor semnificative din domeniu, majoritatea instrumentelor bioinformaticice pentru inferența evolutivă a evenimentelor genomice în cadrul colecțiilor secvențelor genomice de interes continuă să se bazeze pe metode dependente de culturi izolate microbiene și de genomuri de referință de înaltă calitate [11, 16–18]. Depășirea acestor limitări necesită abordări pangenomice bazate pe filogenie, care permit inferența directă a dinamicii de câștig și pierdere a genelor din datele metagenomice. Metodele de reconstrucție a stărilor ancestrale bazate pe procese Markov în timp continuu pot permite recuperarea traiectoriilor fine ale evoluției conținutului genetic de-a lungul liniilor evolutive asociate diferitelor ecosisteme [13, 15, 19, 20]. În acest context, se conturează o nevoie tot mai accentuată pentru dezvoltarea unor metode bioinformaticice eficiente, capabile să integreze modele evolutive dependente de arborii filogenetici cu datele genomice de dimensiuni mari, care acoperă mii de familii de gene derivate din analiza pangenomică. Un astfel de software este esențial pentru detectarea adaptării specifice liniilor evolutive, estimarea ratelor de dinamică genică și delimitarea componentelor conservate față de cele variabile ale pangenomului microbial, reducând astfel un important deficit metodologic în cadrul supravegherii metagenomice. [4, 6, 9, 11, 14, 15, 21–23]. În mod similar, capacitatea de a evalua presiunea selectivă asupra genelor și funcțiilor conservate a devenit esențială pentru identificarea tiparelor genice specifice liniilor evolutive, caracterizate prin retenția, pierderea sau achiziție preferențială de gene, care deviază de la așteptările neutre în contexte ecologice distincte [24, 25].

Scopul și obiectivele lucrării. Teza de doctor are ca scop: (i) elaborarea și validarea unui software bioinformatic modular, scalabil și complet reproductibil destinat reconstrucției și analizei meta-pangenomurilor din date metagenomice urbane; (ii) integrarea unui model probabilistic bazat pe procese Markoviene continue în timp cu metode de inferență filogenetică pentru estimarea

ratelor de dinamică genică și caracterizarea variabilității conținutului genomic; și (iii) dezvoltarea și implementarea unui algoritm statistic pentru identificarea deviațiilor de la așteptările neutre și clasificarea genelor în funcție de regimul selectiv și tiparele evolutive asociate.

Obiective de cercetare propuse:

- Dezvoltarea unui software bioinformatic modular, scalabil și complet reproductibil pentru reconstrucția meta-pangenomului din secvențe genomice etichetate taxonomic, extrase din date metagenomice, care integrează adnotarea structurală și funcțională, inferența grupurilor de gene ortoloage, estimarea parametrilor structurali ai pangenomului, delimitarea fracțiunilor conservate și accesorii, precum și reconstrucția filogenetică, generând rezultate interoperabile și reutilizabile pentru analize evolutive și comparative ulterioare.
- Elaborarea și implementarea unui model probabilistic de tip lanț Markov în timp continuu, formulat pe structura arborelui filogenetic, destinat estimării parametrilor de câștig și pierdere genică și caracterizării dinamicii conținutului genomic de-a lungul liniilor evolutive.
- Dezvoltarea unei metode filogenetice de inferență bazată pe lanțuri Markov în timp continuu, pentru identificarea deviațiilor de la regimul neutru și clasificarea genelor în funcție de regimul selectiv, prin estimarea asimetriei ratelor de câștig și pierdere genică și cuantificarea tendinței evolutive a variației conținutului pangenomic.
- Validarea software-ului dezvoltat, pe întregul flux de analiză, prin evaluarea performanței sale pe seturi de date de referință, aplicarea la date metagenomice urbane pentru reconstrucția meta-pangenomurilor și realizarea inferenței filogenetice bazate pe modele de inferență dezvoltate în vederea estimării evenimentelor de câștig și pierdere genică și clasificării genelor în funcție de regimul de selecție.

Metodologia cercetării științifice. În această teză este dezvoltat un software bioinformatic pentru reconstrucția meta-pangenomurilor din date metagenomice și pentru inferența câștigului și pierderii genice specifice liniilor evolutive, precum și pentru clasificarea genelor ortoloage în funcție de direcția selecției. Aplicabilitatea bioinformatică este demonstrată prin analiza unor seturi empirice de date genomice, o colecție urbană de secvențe metagenomice asamblate pentru genul *Klebsiella* și două colecții de genomuri provenite din culturi izolate (o colecție la nivel de gen *Klebsiella* și un set de date la nivel de specie pentru *K. pneumoniae*), permițând comparații între tipuri de date cu niveluri diferite de calitate și la scări taxonomice distincte. Pentru fiecare set de date a fost reconstruit meta-pangenomul prin predicția și adnotarea secvențelor codificatoare,

gruparea genelor ortoloage și construirea matricelor binare de prezență-absență a genelor ortoloage, cu inferența ulterioară a arborilor filogenetici pe baza alinierii secvențelor genomurilor.

Evoluția conținutului genetic a fost modelată utilizând lanțurile Markov în timp continuu cu două stări (prezență/absență ale genelor ortoloage) pe ramurile arborilor filogenetici. Tranzițiile stărilor de-a lungul ramurilor sunt descrise de matricea de rate $Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$, iar probabilitățile de tranziție pe ramurile filogenetice, cu lungimea t , sunt descrise de $P(t) = e^{Qt}$ [13, 26, 27]. Parametrii de câștig (λ) și pierdere (μ) sunt estimați independent pentru fiecare grup de gene ortoloage prin maximizarea verosimilității și algoritmul Felsenstein [12]. Optimizarea parametrilor este realizată cu optimizatorul quasi-Newton cu constrângeri (L-BFGS-B), implementat în pachetul *stats* din R [28].

Algoritmii PGGL (Pangenome Gene Gain and Loss) și PGGS (Pangenome Gene Selection), fundamentați pe modelele markoviene, analizele statistice și procedurile de vizualizare, au fost implementate în limbajul de programare R [28]. Etapele preliminare ale fluxului de analiză (asamblarea, adnotarea genomică, inferența grupurilor de gene ortoloage, alinierea secvențelor și reconstrucția filogenetică) au fost realizate utilizând instrumente bioinformatiche consacrate [12]. Pentru a asigura reproductibilitatea, portabilitatea și scalabilitatea fluxului de lucru, analizele au fost executate în medii versionate și containerizate, atât pe sisteme locale, cât și pe infrastructuri de calcul de înaltă performanță [29].

Noutatea științifică a rezultatelor cercetărilor. Această teză propune un cadru software pentru analiza evoluției genomice microbiene, bazat pe integrarea meta-pangenomică a datelor metagenomice cu inferența filogenetică. Cadrul permite reconstrucția repertoriilor genice direct din asamblări metagenomice și modelarea dinamicii câștigului și pierderii genice de-a lungul liniilor evolutive. Abordarea este concepută pentru a reflecta structura reală a microbiomului din medii complexe, caracterizate prin asamblări fragmentate, amestecuri de tulpini, transfer orizontal de gene (HGT) și recombinare frecventă. Din punct de vedere metodologic, această lucrare dezvoltă și implementează software-ul PGGL (Pangenome Gene Gain and Loss), un cadru bazat pe verosimilitate maximală și lanțuri Markov în timp continuu, aplicate pe arbori filogenetici, în care fiecare grup ortolog este modelat ca un caracter binar (absent/prezent). Pentru fiecare genă sunt estimate ratele de câștig și pierdere, probabilitățile marginale ale stărilor ancestrale și numărul așteptat de evenimente pe ramuri, calculate prin algoritmul lui Felsenstein, permițând cuantificarea dinamicii genice la nivel de linie evolutivă în cadrul pangenomului. Ulterior, software-ul PGGS (Pangenome Gene Selection) implementează un test pentru evaluarea asimetriei direcționale a procesului de câștig-pierdere genică, unde pentru fiecare genă, un model cu rate egale, care impune

egalitatea ratelor de câștig și pierdere, este comparat cu un model cu rate diferite. Selecția modelului este realizată pe baza criteriului informațional Akaike (AIC), iar genele sunt încadrate în categorii cu predominanță de câștig, predominanță de pierdere sau dinamică simetrică, oferind astfel un indicator cuantificabil al presiunii selective care acționează asupra variației de tip prezență-absență.

Noutatea lucrării constă în extinderea metodelor comparative-filogenetice pe conținutului genic la scară meta-pangenomică, permițând estimarea robustă, la nivel de linie evolutivă, a ratelor de dinamică și a asimetriei direcționale a proceselor de câștig-pierdere genică. În plan aplicativ, software-ul este conceput ca un sistem modular și reproductibil, care integrează gene ortoloage, gene de rezistență la anti-microbiene și factori de virulență într-un cadru filogenetic coerent, fiind orientat către supravegherea genomică de mediu, unde comparabilitatea între seturi de date și între locații reprezintă o cerință metodologică esențială.

Problema științifică de cercetare soluționată. Această teză abordează patru limitări fundamentale ale genomicii evolutive microbiene:

1. Absența unor software și metode validate pentru reconstrucția pangenomurilor microbiene direct din date metagenomice este depășită prin dezvoltarea unui software robust pentru reconstrucția și analiza meta-pangenomurilor.
2. Distorsionarea semnalului filogenetic cauzată de recombinare și transferul orizontal de gene în genomurile microbiene este abordată prin inferența unor filogenii filtrate de semnalul de recombinare, pornind de la date de aliniere a secvențelor.
3. Lipsa unor metode bioinformatică validate pentru cuantificarea dinamicii câștig-pierdere genică la nivel pangenomic este depășită prin dezvoltarea și implementarea unui software dedicat estimării numărului de evenimente și a ratelor de câștig și pierdere genică de-a lungul filogeniilor.
4. Absența unor cadre metodologice pentru estimarea semnalului de selecție la nivel de genă în context pangenomic este abordată prin dezvoltarea unui software implementat în limbajul R, care clasifică genele în funcție de regimul selectiv inferat din datele pangenomice.

Aceste limitări metodologice sunt depășite prin dezvoltarea unui cadru software bazat pe inferență filogenetică și modele Markoviene în timp continuu, care permite reconstrucția repertoriilor genice din asamblări metagenomice și estimarea dinamicii câștig-pierdere genice la nivelul liniilor evolutive. Matricele de prezență-absență a genelor ortoloage sunt analizate, după eliminarea semnalului de recombinare, pe filogenia genomurilor, pentru estimarea ratelor de câștig (λ) și pierdere (μ). Asimetria direcțională este evaluată prin compararea modelului cu rate egale

ale proceselor de câștig și pierdere genică cu un model care permite rate diferite pentru cele două procese, selecția modelului fiind realizată pe baza scorului AIC.

Semnificația teoretică. Această teză dezvoltă teoria genomică comparativă aplicată pangenomurilor microbiene prin elaborarea unui software bioinformatic care permite estimarea directă a evoluției conținutului genic din date metagenomice obținute prin secvențiere de nouă generație. Sunt introduse modele de verosimilitate pentru inferența stărilor de prezență-absență ale genelor pe filogenii, precum și un cadru metodologic pentru evaluarea presiunii de selecție asupra conținutului genic. În mod formal, contribuțiile teoretice ale acestei lucrări sunt următoarele:

- Extinderea analizei datelor metagenomice la nivel pangenomic, prin reconstrucția meta-pangenomurilor din datele de secvențiere și derivarea matricelor binare de prezență-absență ale genelor, urmate de analiza acestora într-un cadru filogenetic comparativ.
- Modelarea prezenței-absenței grupurilor de gene ortoloage ca trăsături discrete, folosind un proces Markov în timp continuu cu două stări definit pe un arbore filogenetic, care permite estimarea ratelor de câștig și pierdere specifice fiecărei gene.
- Extinderea modelului prin reconstrucție ancestrală bazată pe verosimilitate maximală permite inferența stărilor ancestrale ale genelor și estimarea distribuției evenimentelor de câștig și pierdere de-a lungul ramurilor filogenetice, iar integrarea cu filogeniile lipsite de semnale de recombinare permit reducerea *bias*-ului indus de recombinarea omoloagă și corespunzător separare mai robustă între semnalul de moștenire verticală și cel asociat transferului lateral de gene.
- Fundamentarea unei metode bazate pe filogenie pentru modelarea evoluției conținutului genic, prin compararea formală a modelelor simetrice și asimetrice de câștig-pierdere, care furnizează un criteriu statistic pentru distingerea dintre dinamica neutră și procesele selective ce structurează pangenomul de-a lungul istoriei evolutive.

Valoarea aplicativă a lucrării. În termeni practici, această teză furnizează un pachet software bioinformatic destinat supravegherii și monitorizării metagenomice a mediului, transpunând modelarea evolutivă în rezultate direct utilizabile de către cercetători științifici și echipele din domeniul sănătății publice. Pachetul software este conceput pentru a funcționa pe date metagenomice de secvențiere de noua generație și pentru a genera rezultate sumare standardizate și ușor de interpretat, pentru sprijinul procesului decizional în analiza ecosistemelor microbiene complexe. Valoarea sa aplicativă este reflectată în următoarele aplicații operaționale și livrabile:

- Pachetul de instrumente bioinformatică permite reconstrucția meta-pangenomului direct din date metagenomice, facilitând caracterizarea structurii și variabilității repertoriilor genice în

medii dominate de organisme necultivabile sau slab reprezentate genomic, precum apele uzate, aerul și mediile urbane.

- Prin estimarea câștigurilor și pierderilor genice de-a lungul liniilor filogenetice, pachetul bioinformatic furnizează măsuri cantitative ale dinamicii conținutului genomic, care pot fi utilizate pentru identificarea liniilor evolutive cu evoluție rapidă, evaluarea potențialului de adaptare și identificarea țărilor pentru investigații sau intervenții aprofundate.
- Rezultatele generate de software, precum ratele, direcția dinamicii pentru câștig–pierdere ale genelor și numărul de evenimente specifice ramurilor filogenetice, sunt integrate într-un cadru filogenetic comun și livrate sub formă de sumare și vizualizări standardizate, direct utilizabile pentru sisteme de avertizare timpurie, identificarea zonelor de interes și monitorizarea longitudinală pe parcursul campaniilor de eșantionare.
- Software dezvoltat permite analiza integrată a genelor de interes, precum factori RAM, și de virulență în același context evolutiv, facilitând evaluarea riscurilor de mediu și supravegherea coordonată a trăsăturilor cu relevanță directă pentru sănătatea publică.
- Standardizarea procedurilor de inferență permite compararea directă a rezultatelor între locații, intervale temporale și studii independente, susținând analize coerente de supraveghere genomică la diverse nivele geografice.
- Deși este motivat de supravegherea genomică urbană, software dezvoltat este transferabil către alte domenii, inclusiv microbiologia clinică, agricultură, acvacultură, sistemele marine și ecosistemele naturale, fără a necesita ajustare metodologică, susținând analize evolutive oriunde sunt disponibile date metagenomice.
- Prin identificarea liniilor evolutive și a familiilor de gene care prezintă dinamici atipice de câștig–pierdere sau tendințe direcționale, metodele dezvoltate oferă o bază metodologică solidă pentru formularea de ipoteze testabile, care pot fi ulterior investigate prin secvențiere țintită, teste funcționale sau analize epidemiologice.

Tezele științifice înaintate spre susținere:

- Meta-pangenomurile pot fi reconstruite direct din genomuri asamblate din date metagenomice, cu integrarea opțională a microorganismelor necultivabile, generând matrice binare de prezență–absență pentru familiile de gene ortoloage, statistici descriptive ale pangenomului și filogenii pentru secvențelor analizate, potrivite pentru inferențe și analize ulterioare.

- Datele de prezență-absență ale genelor sunt modelate prin procese Markov în timp continuu definite pe arborele filogenetic, ceea ce permite estimarea ratelor de câștig (λ) și pierdere (μ) pe ramuri și cuantificarea acestora la nivel de linie evolutivă.
- Selecția direcțională la nivelul genelor poate fi testată, prin compararea parametrizărilor modelelor Markov în timp continuu, cu rate egale și cu rate distincte, generând statistici interpretabile care cuantifică asimetria dintre achiziție și pierdere.
- Aplicarea metodelor și a software-ului dezvoltate la date metagenomice generează un set reproductibil de rezultate și indicatori (λ , μ , indici de selecție, secvențe prioritizate), produși în cadrul unui flux analitic standardizat și direct utilizabili în analize comparative și în sisteme de avertizare timpurie pentru supravegherea genomică a patogenilor.

Implementarea rezultatelor științifice. Această lucrare a fost transpusă din plan metodologic în practică prin colaborări cu Agenția Națională pentru Sănătate Publică (ANSP), Institutul de Microbiologie și Biotehnologie al Universității Tehnice a Moldovei și Universitatea „Ștefan cel Mare” din Suceava (România).

Aprobarea rezultatelor științifice. Rezultatele fundamentale ale tezei de doctorat au fost prezentate și discutate în cadrul ședințelor și seminarelor Departamentului Inginerie Software și Automatică, Facultatea Calculatoare, Informatică și Microelectronică, Universitatea Tehnică a Moldovei (2022–2025). Acestea au fost raportate și evaluate pozitiv în cadrul a nouă conferințe științifice naționale și internaționale, inclusiv: *International Conference on Nanotechnologies and Biomedical Engineering* (Chișinău, 2025); *International Conference BioGENext: Next Generation Therapy Conference* (Kyiv, 2024); *International Conference on Electronics, Communications and Computing* (Chișinău, 2022, 2024); *Conferința Tehnico-Științifică a Studenților, Masteranzilor și Doctoranzilor* (UTM, Chișinău, 2023).

Publicații la tema tezei. Rezultatele principale ale tezei au fost publicate în 16 lucrări științifice, dintre care 7 articole în reviste internaționale indexate ISI și SCOPUS, incluzând publicații în *Nature Reviews Methods Primers* (IF=56), *Nature Water* (IF=24,24) și *Genome Biology* (IF=9,4), precum și 9 lucrări prezentate și publicate în volumele conferințelor naționale și internaționale (lista publicațiilor este anexată la finalul tezei și al rezumatului tezei de doctorat). Numărul total al publicațiilor este de 49 de lucrări științifice, dintre care 11 indexate ISI și SCOPUS. Autorul are un indice $h = 8$ (SCI Hirsch index), iar numărul total de citări internaționale este >290.

Volumul și structura tezei. Teza cuprinde 116 pagini și include o introducere, 4 capitole, concluzii și recomandări, o bibliografie cu 348 de surse, 5 anexe, 47 de figuri și 11 tabele.

Cuvinte-cheie: bioinformatică, biostatistică, modelare matematică, model Markov în timp continuu, metagenomică, pangenom, biologie computațională, genomică comparativă.

CONȚINUTUL TEZEI

În *Introducere* este argumentată actualitatea și importanța temei de cercetare, prezintă o analiză a nivelului actual al cercetărilor și dezvoltării la subiectul temei, de asemenea sunt expuse scopul și obiectivele tezei, noutatea științifică a rezultatelor obținute, tezele principale înaintate spre susținere, validarea rezultatelor în publicații științifice și lista conferințelor la care au fost expuse și aprobate rezultatele tezei de doctorat.

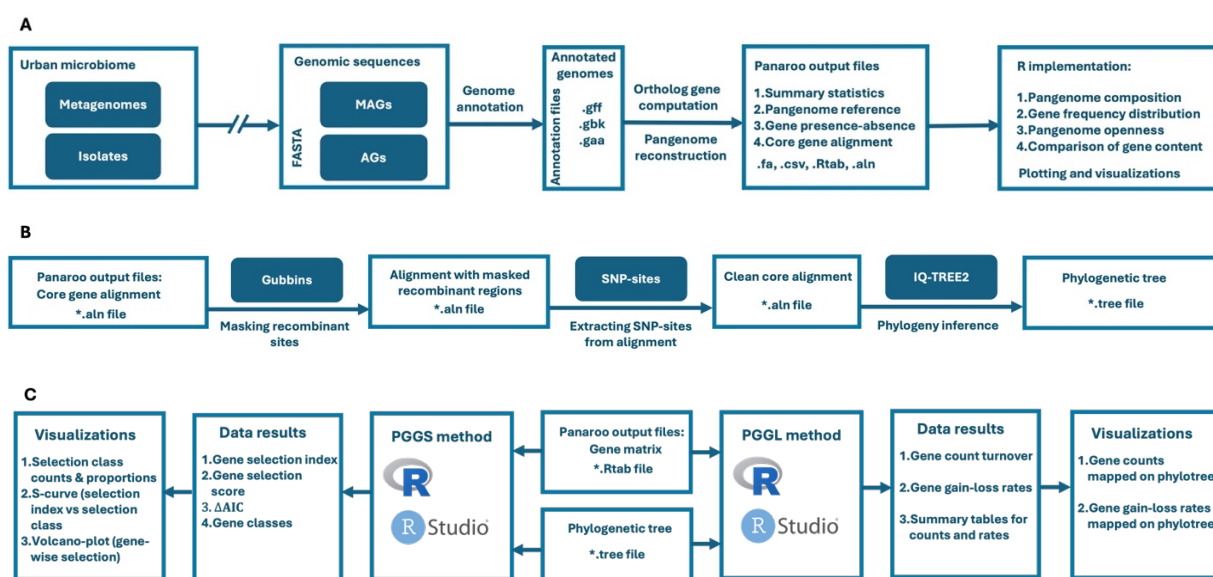


Fig.1. Designul software de bioinformatică pentru reconstrucția și analiza meta-pangenomului. (A) Adnotarea genomică și reconstrucția meta-pangenomului; (B) inferența filogeniei; (C) cuantificarea dinamicii genice (metoda PGGL) și clasificarea genelor în funcție de indicele și scorul de selecție (metoda PGGS).

În *Capitolul 1* este descris contextul microbiomului și metagenomicii urbane, sunt evidențiate limitările analizelor bazate pe secvențele obținute prin metode de cultură a microorganismelor și este prezentată abordarea meta-pangenomică ca metodă necesară pentru recuperarea diversității genelor accesorii care guvernează adaptarea ecologică și riscul sănătății publice. Sunt definite componentele genomice conservate și cele accesorii, gradul de deschidere al pangenomului și motivația utilizării inferenței bazate pe filogenie, fiind stabilit cadrul conceptual pentru modelele dezvoltate în capitolele următoare.

În *Capitolul 2* sunt caracterizate și prezentate sursele și specificul datelor utilizate, precum și rezultatele software-ului bioinformatic dezvoltat. Acesta realizează adnotarea structurală și

funcțională a secvențelor, efectuează gruparea ortologică a genelor, inferența filogeniei cu eliminarea regiunilor recombinate (Figura 1B) și realizează calculul statisticilor pentru pangenom (Figura 1A). Sunt introduse și implementate în limbajul R două software de modelare bazate pe verosimilitatea maximală și lanțurile Markov în timp continuu: (1) PGGL (*Pangenome Gene Gain Loss*), care cuantifică, estimează ratele și cartează evenimentele de câștig și pierdere ale genelor pe arborii filogenetic și generează sumare pentru ratele de câștig și pierdere la nivel de ramuri filogenetice sau grupuri taxonomice de interes; și (2) PGGS (*Pangenome Gene Selection*), compară asimetria dintre ratele de câștig și pierdere, pentru a detecta tendințele evolutive și clasifică genele în clase în baza acestor tendințe (Figura 1C).

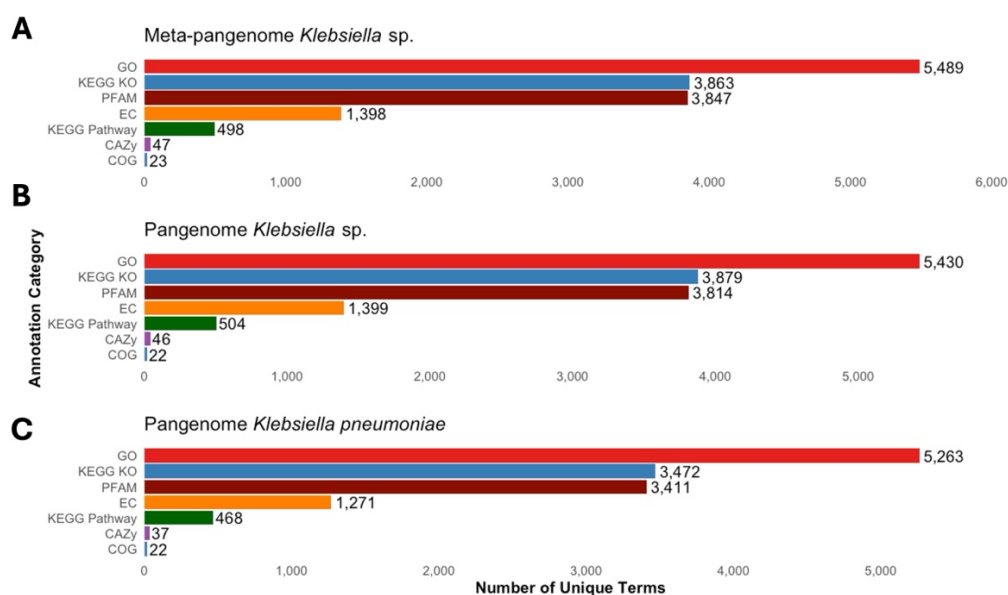


Fig.2. Sumarul adnotării funcționale pentru (A) meta-pangenomul genului *Klebsiella* , (B) pangenomul genului *Klebsiella* și (C) pangenomul pentru *Klebsiella pneumoniae*.

Capitolul 3 prezintă rezultatele cercetării privind validarea software-ului bioinformatic pentru reconstruirea și analiza meta-pangenomurilor atât din genomuri asamblate din date metagenomice, cât și din genomuri izolate. Fluxul software efectuează identificarea genelor, adnotarea structurală și funcțională, gruparea genelor după ortologie, crearea sumarelor pentru frecvența genelor, estimarea deschiderii pangenomurilor și diverse metode de vizualizarea în cadrul unui protocol analitic unificat. Intrările și ieșirile sunt armonizate pentru a permite comparații directe și controlate între diferite contexte ecologice și scări taxonomice.

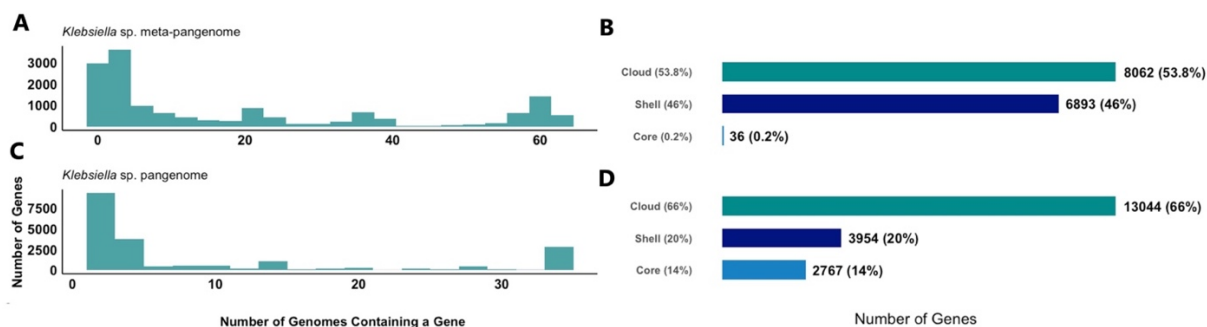


Fig.3. Distribuția comparativă a frecvenței genelor (A, C) și compoziția genică (core, shell și cloud) în meta-pangenom (A, B) și pangenom (C, D).

Rezultatele adnotării structurale evidențiază diferențe importante între sursele de date. Genomurile din culturi prezintă în mod constant un număr mai mare însă mai puțin variabil de secvențe codificatoare, secvențe ARNt aproape complete, dimensiuni genomice mai mari, precum și un număr mai redus de fragmente genomice asamblate comparativ cu genomurile asamblate din datele metagenomice. Aceste paterne reflectă limitări bine cunoscute ale metagenomicii bazate pe citiri scurte, inclusiv fragmentarea, acoperirea neuniformă și erorile de grupare taxonomică [16, 17, 30, 31]. Comparările directe între seturile de date, împreună cu analize exploratorii ale conținutului genetic, evidențiază același tipare structurale, unde genomurile obținute din culturi prezintă o variabilitate redusă a conținutului de gene și o structură genomică mai omogenă, în timp ce genomurile asamblate din date metagenomice manifestă o dispersie mai mare a profilurilor de gene, reflectând eșantionarea eterogenă și gradul variabil de completitudine al asamblărilor metagenomice [16, 17, 30, 31].

Profilarea funcțională (GO, KEGG, PFAM, EC, COG, CAZy) permite identificarea funcțiilor conservate (Figura 2). Speciile reprezentate prin genomuri cultivate de înaltă calitate prezintă o variabilitate redusă a termenilor ontologici și a domeniilor proteice, în concordanță cu adnotările uniforme și asamblările genomice aproape complete (Figurile 2B–C) [32, 33].

În contrast, meta-pangenomurile derivate din date metagenomice recuperează un repertoriu funcțional mai larg, în special la nivelul claselor enzimatice și al căilor metabolice, reflectând atât amplitudinea ecologică, cât și capacitatea metagenomicii de a surprinde diversitatea necultivată prezentă în diverse nișe ecologice (Figura 2A) [16, 17]. Genomurile obținute din izolate cultivate de *K. pneumoniae* ating cea mai înaltă fidelitatea a adnotării și cea mai scăzută variație intra-specifică, păstrând totodată o acoperire extinsă a claselor enzimatice, termenilor ontologici și domeniilor proteice. În ansamblu, aceste rezultate susțin o perspectivă hibridă, în care genomurile asamblate din date metagenomice extind dimensiunea ontologică funcțională, în timp ce genomurile obținute

din culturi sporesc încrederea în inferențele la nivel structural al secvențelor și reduc artefactele introduse de fragmentarea asamblărilor.

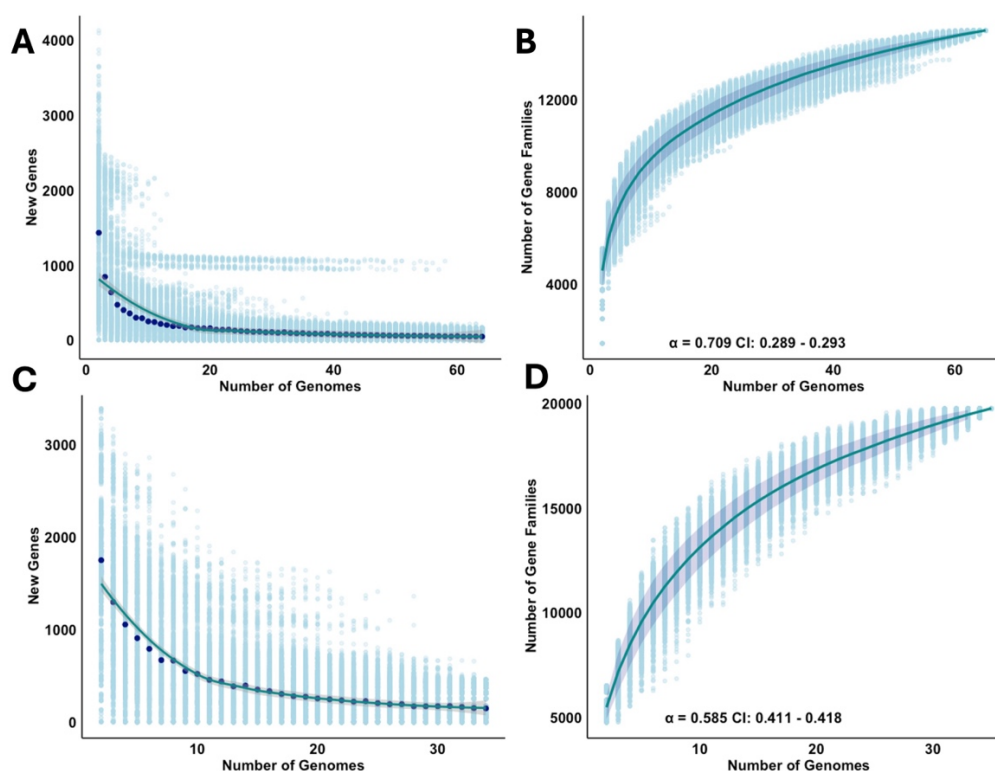


Fig.4. Dinamica descoperirii genelor (A, C) și numărul cumulativ de familii de gene unice ajustat prin modelul legii lui Heaps (B, D) în meta-pangenom (A, B) și pangenom (C, D).

Clasificarea genelor după ortologie și în funcție de frecvența în genomuri recuperează arhitectura clasică *core-shell-cloud* (corespund prezenței genelor în >95%/10-95%/<10% genomuri) a pangenomurilor bacteriene (Figura 3) [34, 35]. Meta-pangenomul construit exclusiv din genomuri asamblate din date metagenomice prezintă cel mai redus *core* (genom comun) și cea mai extinsă componentă de tip *shell*, reflectând prezența unui număr mare de gene cu frecvență scăzută, caracteristic datelor din medii ecologice diverse (Figura 3A–B). Setul mono-specific de *K. pneumoniae* se distinge prin cel mai extins *core* și o componentă *cloud* pronunțată, în concordanță cu o amplitudine filogenetică mai restrânsă și o eșantionare clinică mai uniformă, în timp ce datele obținute din culturi pentru speciile din genul *Klebsiella* ocupă o poziție intermediară (Figura 3C–D). Histogramele frecvenței genelor evidențiază distribuția asimetrică în formă de “U”, cu majoritatea familiilor de gene, fie comune, fie rare și un număr redus de familii cu prevalență intermediară, conform teoriei clasice a pangenomului (Figura 3B, D) [7, 34, 35].

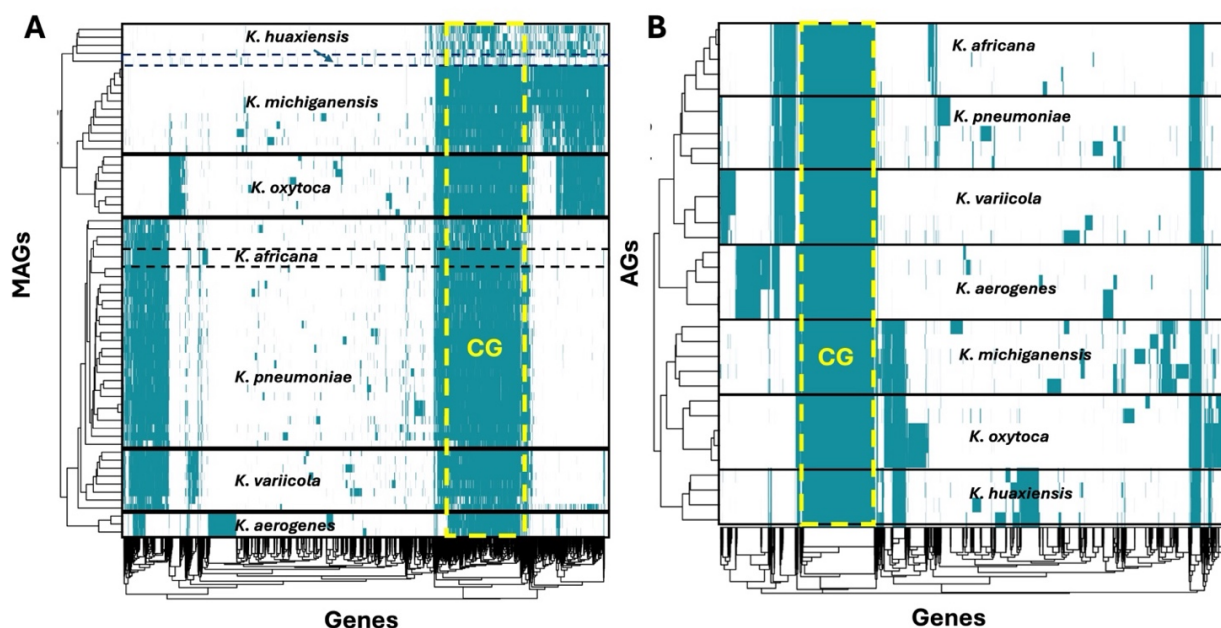


Fig.5. Clasificarea ierarhică comparativă a genelor și genomurilor în meta-pangenom (A) și pangenom (B). Notă: CG-genom comun, MAGs-genomuri asamblate din date metagenomice, AGs-genomuri asamblate.

Dinamica descoperirii de gene noi în genomuri a fost cuantificată prin analize de rarefiere și prin ajustarea numărului cumulativ de familii de gene la Legea lui Heaps (Figura 4) [7, 36, 37]. Pangenomul bazat pe culturi rămâne pronunțat deschis pentru toate permutările de eșantionare, indicând o descoperire continuă de familii noi de gene (Figura 4C–D). Meta-pangenomul derivat din date metagenomice este de asemenea deschis, însă prezintă o rată de creștere mai lentă, sugerând o saturație parțială a funcțiilor la nivel de gen (Figura 4A–B).

Clasificarea ierarhică a genelor din matricea de prezență-absență evidențiază nuclee dense, aproape universale, peste care se suprapun „insule” de gene accesorii (Figura 5). În meta-pangenom, modulele accesorii apar fragmentate și restrânse pe clade, formând insule discontinue. Coloanele cu semnale rare corespund genomurilor cu dimensiuni scurte și fragmentate, un patern caracteristic genomurilor recuperate incomplet din date metagenomice (Figura 5A) [16, 17, 30, 31]. În pangenomul de culturi multi-specie, genomurile se separă clar și recapitulează taxonomia speciilor (Figura 5B). Analiza componentelor principale a conținutului binar de gene evidențiază o dispersie largă pentru genomurile recuperate din metagenom, reflectând eterogenitatea ecologică și compoziția de specii, în timp ce speciile bazate pe genomi cultivați formează cluster compacte și bine separate, în concordanță cu taxonomia. În cadrul speciei *K. pneumoniae*, principalele tipuri de

secvență ocupă regiuni distincte, relevând repertorii accesorii specifice liniilor evolutive, în acord cu analizele de genomică populațională ale speciei [38, 39].

Contribuția principală a acestui capitol constă în validarea instrumentelor bioinformaticice pentru reconstrucția pangenomului din date metagenomice pe cele trei seturi de date: (1) meta-pangenomul bazat pe genomuri asamblate din date metagenomice maximizează amplitudinea ecologică și diversitatea componentelor accesorii; (2) genomurile obținute din culturi multi-specie oferă rezoluție la nivel de gen, cu o structură bine definită la nivel de specie și un nivel redus de zgomot genetic; (3) datele clinice mono-specie permit rezolvarea arhitecturii liniilor clonale, caracterizată printr-un nucleu aproape universal și module discrete specifice tipurilor de secvență. Astfel, strategia de eșantionare, calitatea asamblării și rezoluția filogenetică acționează împreună în determinarea structurii pangenomului, a gradului său de deschidere și a vizibilității componentelor accesorii. Aceste rezultate constituie o bază solidă pentru modelarea evolutivă dezvoltată ulterior în teză, permițând aplicarea metodelor de inferență bazate pe filogenii.

În **Capitolul 4** sunt dezvoltate și implementate două instrumente bioinformaticice complementare, bazate pe modele Markoviene în timp continuu și filogenie. Metoda PGGL (**Pangenome Gene Gain and Loss**) modelează prezența/absența fiecărui gene ca o trăsătură binară care evoluează de-a lungul unui arbore filogenetic al speciilor, într-un cadru Markov în timp continuu. Metoda estimează ratele de câștig (λ) și pierdere (μ) ale genelor și reconstruiește probabilistic dinamica acestor evenimente de-a lungul ramurilor arborelui filogenetic, prin inferența stărilor ancestrale și estimarea tranzițiilor la scară pangenomică (Algoritmele 1 și 2).

Algoritmul 1. Inferența fluxului și estimarea ratelor la nivel de genă (PGGL-Gene).

Intrare: Un arbore filogenetic înrădăcinat, strict bifurcat, cu lungimi ale ramurilor; un vector binar de prezență-absență pentru gena g , aliniat la vârfurile arborelui; specificarea modelului simetric sau cu rate diferite; informația <i>a priori</i> la rădăcina arborelui (uniformă sau staționară); și un prag δ pentru apelarea evenimentelor.
Ieșire: Pentru fiecare ramură a arborelui, evenimentele inferate de câștig și pierdere ale genei, împreună cu estimările de maximă verosimilitate ale ratelor ($\hat{\lambda}$) de câștig și pierdere ($\hat{\mu}$) pentru gena g .
Procedură: 1. Specificarea modelului: Evoluția prezenței-absenței genei este modelată ca un proces Markov în timp continuu cu două stări, definite pe arborele filogenetic, cu tranziții corespunzătoare achiziției genei ($0 \rightarrow 1$, rată λ) și deleției genei ($1 \rightarrow 0$, rată μ). 2. Calculul verosimilității și estimarea ratelor: Stările observate la vârfuri sunt codificate ca vectori de verosimilitate, iar verosimilitatea datelor este calculată descendent utilizând algoritmul lui Felsenstein. Ratele specifice de câștig și pierdere ($\hat{\lambda}, \hat{\mu}$) a genei sunt estimate prin verosimilitate maximă, fie sub parametrizarea cu rate egale, fie sub parametrizarea cu rate diferite. 3. Reconstrucția stărilor ancestrale: Ratele estimate sunt utilizate la calcularea probabilităților posterioare marginale pentru prezența genei în toate nodurile interne, prin propagarea ascendentă pe arbore. 4. Apelarea evenimentelor pe ramuri: Pentru fiecare ramură, sunt comparate probabilitățile posterioare ale prezenței genei între nodul părinte și nodul descendent. Un eveniment de câștig sau pierdere este înregistrat atunci când diferența posterioară depășește pragul predefinit δ. 5. Raportare: Pentru fiecare ramură sunt raportate evenimentele inferate de câștig și pierdere, împreună cu estimările corespunzătoare ale ratelor la nivel de genă.

Metoda PGGGS (*Pangenome Gene Selection*) implementată ca software în limbajul R, testează la nivel de genă, dacă dinamica câștig–pierdere este compatibilă cu un regim neutru sau indică o preferențial pentru câștig sau pierdere, cuantificând intensitatea și direcția prin mărimea efectului definit pe baza raportului dintre ratele de câștig și pierdere. Suportul pentru modelele alternative este evaluat utilizând Criteriul Informațional Akaike (Algoritmele 1 și 3) [40, 41]. Aplicat atât pe metagenomurile urbane cu amplitudine ecologică largă, cât și seturilor de date bazate pe microorganisme cultivate, inclusiv pe genomurile clinice, acest instrument bioinformatic evidențiază tiparele de câștig și pierdere a genelor structurate pe liniile evolutive și permite clasificarea genelor în regimuri neutre sau cu dinamică orientată către câștig sau pierdere, în concordanță cu presiunile selective de retenție, câștig sau pierdere [5, 6, 36, 42, 43].

Distribuțiile normalizate ale ratelor de câștig (λ) și pierdere (μ) evidențiază un regim consistent dominat de pierdere la nivel de specie, cu $\mu > \lambda$ atât în pangenomurile bazate pe izolate, cât și în meta-pangenomuri, în acord cu tendința reducere a genomurilor cunoscută la bacterii (Figura 6) [6, 42].

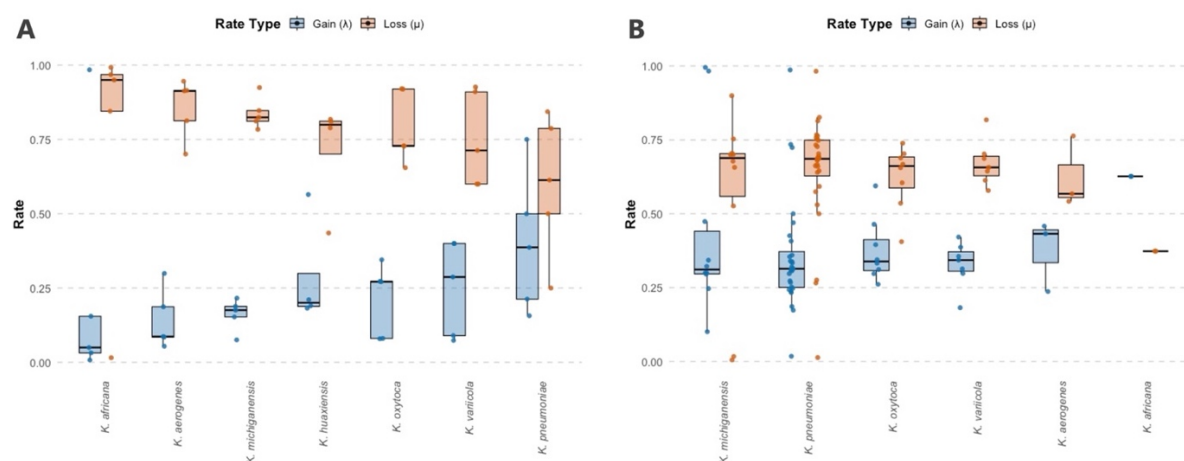


Fig.6. Distribuția sumară, normalizată, a ratelor de câștig și pierdere a genelor la nivel de specie în (A) pangenom și (B) meta-pangenom.

Meta-pangenomul prezintă distribuții mai largi și puternic asimetrice ale ratelor, indicând o dinamică eterogenă a conținutului genetic, în care majoritatea genomurilor evoluează lent, iar un subset restrâns prezintă rate crescute de pierdere, însoțite de creșteri localizate și specifice ale ratelor de câștig, în concordanță cu oportunități ecologice episodice și cu acces variabil la bazine de gene mobile (Figura 6B). În contrast, seturile de date bazate pe izolate prezintă distribuții mai concentrate și cea mai pronunțată tendință către pierdere, reflectând calitatea superioară ale asamblărilor (Figura 6A) [17, 31].

Algoritmul 2. Apelarea evenimentelor de câștig-pierdere a genelor la nivel de pangenom (PGGL)

Intrare: Un arbore filogenetic înrădăcinat, strict bifurcat, cu lungimi ale ramurilor; o matrice binară de prezență-absență P (genomuri $\times$ gene) aliniată la vârfurile arborelui; un model de câștig-pierdere; informații *a priori* pentru rădăcină; și un prag δ pentru apelarea evenimentelor.

Ieșire: Pentru fiecare genă și fiecare ramură, evenimentele estimate de câștig și pierdere ale genelor, împreună cu estimările prin verosimilitatea maximală specifice genei pentru ratele de câștig și pierdere ($\hat{\lambda}_g, \hat{\mu}_g$). Sunt generate sumare la nivel de genom pentru câștigurile și pierderile de gene obținute prin agregarea evenimentelor pentru toate genele de-a lungul traiectoriei filogenetice a fiecărui genom.

Procedură:

1. **Alinierea matrice-arbore:** Se verifică faptul că rândurile matricei P (genomurile) corespund ordinii vârfurilor din arborele filogenetic.
2. **Inferența la nivel de genă:** Fiecare coloană din P este tratată ca un vector de prezență-absență specific genei x_g . Genele invariabile sunt excluse, iar pentru fiecare genă rămasă se aplică **Algoritmul 1 (PGGL-Gene)** pentru a estima evenimentele de câștig și pierdere la nivel de ramură și pentru a estima ratele ($\hat{\lambda}_g, \hat{\mu}_g$).
3. **Agregarea la nivel de pangenom:** Toate estimările obținute la nivel de genă sunt combinate într-un tabel unic de evenimente la nivelul întregului pangenom.
4. **Sumare la nivel de genom:** Pentru fiecare genom (vârf), evenimentele inferate pentru toate genele sunt agregate de-a lungul traiectoriei sale filogenetice de la rădăcină la vârf, pentru a calcula numărul și ratele specifice genomului de câștig și pierdere a genelor.
5. **Raportare:** Sunt returnate sumare pentru dinamica pangenomului la nivel de genă, ramură și genom.

Metoda PGGS (*Pangenome Gene Selection*) implementată în limbajul R, utilizează ratele de verosimilitate maximală, estimate la nivel de genă, pentru a discrimina între dinamica simetrică și direcționată ale genelor.

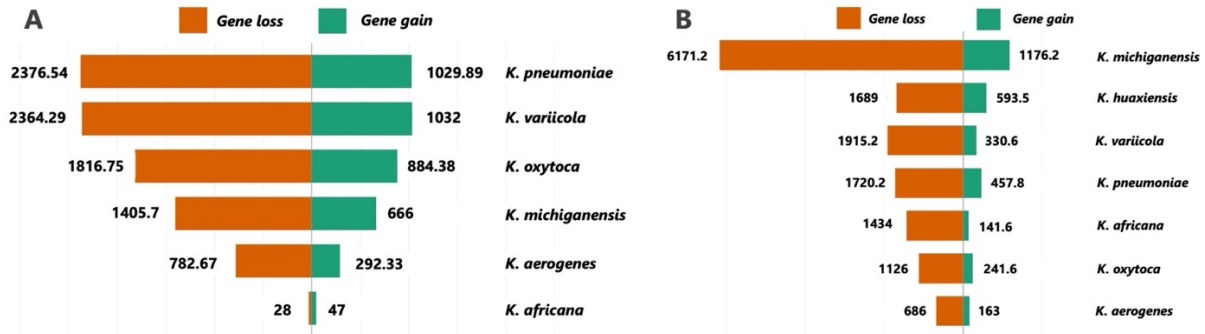


Fig.7. Câștigurile și pierderile de gene la nivel de specie în (A) meta-pangenom și (B) pangenom.

Pentru fiecare genă, algoritmul PGGS compară un regim cu rate egale ($\lambda = \mu$), compatibil cu o dinamică nedirecționată în condiții de selecție aproape neutră sau fluctuantă, cu un regim cu rate diferite ($\lambda \neq \mu$), care surprinde procese direcționale atribuibile selecției pe termen lung sau asimetriilor de proces, precum tendința de achiziție sau pierdere a genelor și aportul inegal de transfer orizontal de gene (Algoritmii 1 și 3) [6, 42, 43, 46, 47]. Sumarele la nivel de genom sau specie ale evenimentelor absolute de câștig și pierdere a genelor evidențiază suplimentar predominanța pierderilor față de câștiguri de gene, cu o variație considerabilă între specii (Figura 7).

Câștigurile sunt concentrate într-un subset de linii evolutive, fiind compatibile cu achiziția de gene din bazele de gene accesorii asociate mobilității genetice, mecanismelor de apărare, metabolismului adaptat nișei și rezistenței anti-microbiene (Figura 7) [43–45].

Algoritmul 3. Algoritmul de inferență a selecției direcționale asupra genelor (PGGS)

Intrare: O filogenie înrădăcinată, strict bifurcată, cu lungimi ale ramurilor; o matrice binară de prezență-absență P (genome × gene) aliniată la vârfurile arborelui; un interval de frecvență care definește genele informative; un model de câștig–pierdere (rate egale sau diferite); o distribuție <i>a priori</i> la rădăcină; precum și un prag pentru evenimente.
Ieșire: Pentru fiecare genă informativă, estimări ale ratelor de câștig și pierdere specifice modelului, statistici de comparare a modelelor (<i>log-verosimilitatea</i>, <i>AIC</i>, <i>ΔAIC</i>) un scor de selecție la nivel de genă și clasa de selecție atribuită.
Procedură: Filtrarea genelor: Analiza este restrânsă la genele informative, prin excluderea genelor invariante, sau extrem de rare și frecvente, pe baza frecvenței acestora în pangenom. Ajustarea modelelor: Pentru fiecare genă păstrată, se aplică de două ori Algoritmul 1 (PGGL-Gene), o dată sub modelul cu rate egale și o dată sub modelul cu rate diferite, pentru a estima ratele de câștig și pierdere. Compararea modelelor: Asimetria ratelor este cuantificată prin compararea modelelor utilizând criterii informaționale (<i>AIC</i> și <i>ΔAIC</i>). Estimarea scorului de selecție: Se derivă un indice de selecție la nivel de genă pe baza mărimii relative și a asimetriei ratelor de câștig și pierdere inferate. Clasificarea genelor: Fiecare genă este atribuită unei clase discrete de selecție (selecție neutră, slabă, moderată și puternică), pe baza scorului de selecție și suportului modelului. Raportarea: Se returnează un tabel la nivel de genă care sumează estimările ratelor, statisticile de comparare a modelelor, scorurile de selecție și clasele de selecție.

Suportul modelului este evaluat folosind scorul AIC [40, 41]. Direcția și magnitudinea efectului sunt estimate prin indicele de selecție și scorul de selecție, ambele definite ca funcții ale raportului dintre rata de câștig (λ) și rata de pierdere (μ) ale genelor. Prin utilizarea rapoartelor, aceste mărimi rămân independente de scalarea absolută a ratelor și comparabile între seturi de date cu lungimi diferite ale ramurilor filogenetice sau calibrări temporale distincte. Etichetele de clasă oferă o stratificare discretă, direct utilizabilă pentru analiza comparativă și ierarhizarea genelor. Aplicarea metodei PGGS detectează un semnal de selecție direcțională pronunțat în datele de tip meta-pangenom, provenite din medii ecologic diverse, caracterizat prin predominanța genelor supuse selecției care favorizează pierderea, alături de o fracțiune mai redusă, dar substanțială, de gene aflate sub selecție favorizantă câștigului de gene (Figura 8A). Acest semnal de selecție direcțională este atenuat în pangenomurile obținute din culturi la scară de gen (*Klebsiella* sp.) și tinde către simetrie efectivă în genomurile cultivate la rezoluția de specie (*K. pneumoniae*), în concordanță cu ipoteza conform căreia lărgirea nișei ecologice și eterogenitatea sursei de gene intensifică selecția direcțională asupra conținutului genomic (Figura 8B) [36, 48, 49]. Clasificările rămân robuste față de alegerea distribuției *a priori* la rădăcină (uniformă sau staționară) și față de perturbările moderate ale lungimilor ramurilor, precum și pentru setărilor de optimizare, iar ierarhizările bazate pe *ΔAIC* pentru genele puternic asimetrice sunt reproductibile între utilizările repetate a metodei. Împreună,

1. *Journal of the American Medical Association*, 1997; 278: 1039-1044.

această teză ca nucleul inferențial al metodelor PGGL și PGGS. Implementarea PGGL asigură acoperire completă a nodurilor și o acuratețe efectivă de aproximativ 96%, apropiată de cea obținută prin metoda SCM Bayesiană, menținând în același timp rate de eroare echilibrate (Tabelul 1). În contrast, parcimonia Fitch generează o proporție substanțială de noduri interne ambigue, ceea ce conduce la o reducere semnificativă a acurateței efective atunci când stările nerezolvate sunt incluse în evaluare (Tabelul 1). Aceste rezultate demonstrează că abordarea bazată pe verosimilitatea maximală, care stă la baza PGGL și PGGS, oferă o inferență precisă și eficientă din punct de vedere computațional a stărilor ancestrale, comparabilă cu metoda SCM Bayesiană, dar considerabil mai scalabilă pentru filogenii de mari dimensiuni.

Tabelul 1. Metrice comparative de performanță pentru metodele de reconstrucție a stărilor ancestrale

Metodă	Acuratețe	Precizie	Sensibilitate	Specificitate	Noduri ambigue
Verosimilitate Maximală	0.963	0.966	0.973	0.945	0
Bayes (SCM)	0.963	0.964	0.975	0.943	0
Parcimonie (Fitch)	0.944	0.964	0.944	0.943	392

Spre deosebire de abordările euristice bazate exclusiv pe frecvența genelor (core/shell/cloud), care descriu variabilitatea, dar nu pot distinge între câștigurile și pierderile recente pe filogenii, și pierderile ancestrale partajate de-a lungul evoluției, cadrul metodologic propus este explicit dependent de filogenie și ancorează dinamica câștigului și pierderii genelor în structura evolutivă a arborelui pe care aceste evenimente s-au acumulat. Rezultatele generate sunt direct utilizabile pentru reprezentări grafice și analize de adnotare funcțională și pot fi generalizate ușor între diferite tipuri și surse de date de secvențiere, furnizând o bază unitară pentru inferențe comparative pentru studiul pangenomului.

Fiecare capitol al tezei se încheie cu o sinteză critică a principalelor rezultate obținute, iar capitolul final de concluzii și recomandări integrează rezultatele esențiale, diseminate prin publicații științifice, evidențiind relevanța teoretică și valoarea aplicativă a cercetărilor privind analiza evolutivă a datelor metagenomice.

CONCLUZII GENERALE ȘI RECOMANDĂRI

Această teză demonstrează că metagenomica și pangenomica pot fi integrate într-un cadru analitic unificat, dependent de filogenie, care permite elucidarea modului în care conținutul genic

este organizat și se modifică. Avansul conceptual esențial constă în tratarea prezenței–absenței genelor ca trăsătură evolutivă pe o filogenie fixă, ceea ce face posibilă localizarea evenimentelor, estimarea ratelor și testarea preferențial al fluxului genic, între diferite surse de asamblări genomice și rezoluții taxonomice. În consecință, principalele concluzii sunt:

1. A fost dezvoltat un cadru software bioinformatic robust și complet reproductibil pentru reconstrucția și analiza meta-pangenomurilor, care permite integrarea genomurilor cultivate și date metagenomice eterogene într-o reprezentare unificată a conținutului genic, permițând analize genomice comparative și evolutive.
2. A fost dezvoltată și integrată o metodă robustă de reconstrucție filogenetică, bazată pe eliminarea semnalelor de recombinare și pe alinierea secvențelor conservate la nivel de pangenom, care permite obținerea unui cadru evolutiv explicit necesar modelării probabilistice a evoluției prezenței–absenței genelor și estimării proceselor de câștig și pierdere genică în colecții eterogene de genomuri.
3. A fost dezvoltat și implementat în limbajul R un algoritm de inferență a câștigului și pierderii de gene bazat pe metoda verosimilității maxime și filogenie (algoritmul PGGL), care modelează evoluția stărilor de prezență–absență ale genelor ca proces Markov în timp continuu, permițând estimarea cantitativă a ratelor și a numărului de evenimente de câștig și pierdere genică pe ramurile filogenetice.
4. A fost elaborat și implementat în R un algoritm de clasificare și inferență a presiunii selective asupra genelor (PGGS), bazat pe compararea explicită a ratelor de câștig și pierdere genică estimate sub modele evolutive simetrice și asimetrice, care permite identificarea regimurilor evolutive cu tendință direcțional de câștig sau pierdere, susținute statistic, la nivel de genă și linie evolutivă.
5. Cadrul de reconstrucție a meta-pangenomului, împreună cu algoritmi de inferență PGGL și PGGS, au fost validați empiric pe seturi de date provenite din genomuri cultivate și genomuri asamblate din date metagenomice, la diferite nivele taxonomice, generând reprezentări coerente pentru prezența–absența genelor, inferențe stabile pentru proceselor de câștig și pierdere genică și clasificări consistente ale genelor după tendințele direcționale a selecției.
6. Metoda dezvoltată, bazată pe inferență prin verosimilitate maximală în cadrul unui model Markov în timp continuu, a fost evaluată comparativ la nivelul reconstrucției stărilor ancestrale ale genelor față de metoda de maparea stocastică Bayesiană și parcimonia Fitch, demonstrând o acuratețe fiabilă a reconstrucției, cu ambiguitate redusă și eficiență computațională superioară în analize la scară genomică.

Pentru a conferi caracter durabil acestor rezultate, pangenomica metagenomică trebuie structurată ca un cadru analitic integrat, conform principiilor FAIR (identificabilitate, accesibilitate, interoperabilitate și reutilizabilitate a datelor), cu control riguros al calității și integrarea explicită a etapelor de asamblare, inferență evolutivă și raportare pentru sănătatea publică:

1. Se recomandă ca analizele să fie realizate concomitent la nivelul datelor de mediu, al genomurilor izolate și al celor clinice, utilizând praguri analitice uniforme și un model de date comun, astfel încât semnalele inferate să fie direct comparabile între contexte distincte și de-a lungul seriilor longitudinale.
2. Pentru seturile de date cu valoare informațională ridicată, în special pentru genomurile asamblate din date metagenomice care conțin regiuni repetitive, se recomandă utilizarea secvențierii cu citiri lungi sau a asamblărilor hibride. Recuperarea completă a operonilor precum și a regiunilor mobile bogate în repetiții, vor reduce absențele genice generate de fragmentările asamblărilor și vor permite delimitarea mai precisă a insulelor accesorii, precum și vor îmbunătăți localizarea evenimentelor de câștig–pierdere a genelor.
3. Se recomandă ca instrumentele de mapare a evenimentelor de câștig-pierdere a genelor și de analiză a direcției fluxului genic să fie dezvoltate ca pachet software în R (bibliotecă R), cu interfețe stabile, seturi demonstrative, documentație reproductibilă și funcții dedicate de vizualizare.
4. Se recomandă evaluarea adițională a inferenței pentru robustețe prin analize de sensibilitate față de alegerea și localizarea rădăcinii arborelui filogenetic, scalarea lungimilor ramurilor și setările de optimizare, în special pentru genele care prezintă asimetrie direcțională pronunțată. O astfel de evaluare consolidează interpretarea biologică a indicatorilor de selecție și reduce riscul unor concluzii influențate predominant de parametrizarea tehnică.
5. Integrarea inferențelor de câștig–pierdere genică cu adnotări funcționale standardizate ale secvențelor (de exemplu, COG, KEGG, PFAM), esențiale pentru interpretarea biologică a paternelor observate și pentru realizarea unor analize de interpretare funcțională comparabile între seturi de date și contexte ecologice distincte.

Bibliografie

1. Koonin, Eugene V., and Yuri I. Wolf. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* 36: 6688–6719. <https://doi.org/10.1093/nar/gkn668>.
2. Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405. Nature Publishing Group: 299–304. <https://doi.org/10.1038/35012500>.
3. Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W. J. van Passel, and Adam Eyre-Walker. 2015. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in Microbiology* 23: 598–605. <https://doi.org/10.1016/j.tim.2015.07.006>.
4. Treangen, Todd J., and Eduardo P. C. Rocha. 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLOS Genetics* 7. Public Library of Science: e1001284. <https://doi.org/10.1371/journal.pgen.1001284>.
5. Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development* 15. Genomes and Evolution: 589–594. <https://doi.org/10.1016/j.gde.2005.09.006>.
6. McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017. Why prokaryotes have pangenomes. *Nature Microbiology* 2. Nature Publishing Group: 1–5. <https://doi.org/10.1038/nmicrobiol.2017.40>.
7. Vernikos, George, Duccio Medini, David R Riley, and Hervé Tettelin. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23. Host–Microbe Interactions: Bacteria • Genomics: 148–154. <https://doi.org/10.1016/j.mib.2014.11.016>.
8. Ma, Bing, Michael France, and Jacques Ravel. 2020. Meta-Pangenome: At the Crossroad of Pangenomics and Metagenomics. In *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, ed. Hervé Tettelin and Duccio Medini, 205–218. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_9.
9. Danko, David, Daniela Bezdan, Evan E. Afshin, Sofia Ahsanuddin, Chandrima Bhattacharya, Daniel J. Butler, Kern Rei Chng, et al. 2021. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* 184: 3376–3393.e17. <https://doi.org/10.1016/j.cell.2021.05.002>.
10. Boucher, Yan, Christophe J. Douady, R. Thane Papke, David A. Walsh, Mary Ellen R. Boudreau, Camilla L. Nesbø, Rebecca J. Case, and W. Ford Doolittle. 2003. Lateral Gene Transfer and the Origins of Prokaryotic Groups. *Annual Review of Genetics* 37. Annual Reviews: 283–328. <https://doi.org/10.1146/annurev.genet.37.050503.084247>.
11. Zolfo, Moreno, Francesco Asnicar, Paolo Manghi, Edoardo Pasolli, Adrian Tett, and Nicola Segata. 2018. Profiling microbial strains in urban environments using metagenomic sequencing data. *Biology Direct* 13: 9. <https://doi.org/10.1186/s13062-018-0211-z>.
12. Liu, Shaopeng, Judith S. Rodriguez, Viorel Munteanu, Cynthia Ronkowski, Nitesh Kumar Sharma, Mohammed Alser, Francesco Andreace, et al. 2025. Analysis of metagenomic data. *Nature Reviews Methods Primers* 5. Nature Publishing Group: 1–28. <https://doi.org/10.1038/s43586-024-00376-6>.
13. Felsenstein, Joseph. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17: 368–376. <https://doi.org/10.1007/BF01734359>.
14. Pagel, Mark. 1999. The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. *Systematic Biology* 48. [Oxford University Press, Society of Systematic Biologists]: 612–622.
15. Hendriksen, Rene S., Patrick Munk, Patrick Njage, Bram van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, et al. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications* 10. Nature Publishing Group: 1124. <https://doi.org/10.1038/s41467-019-08853-3>.
16. Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35. Nature Publishing Group: 833–844. <https://doi.org/10.1038/nbt.3935>.
17. Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. Minimum information about a single amplified genome

- (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35. Nature Publishing Group: 725–731. <https://doi.org/10.1038/nbt.3893>.
18. Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* 31. Nature Publishing Group: 533–538. <https://doi.org/10.1038/nbt.2579>.
 19. Ishikawa, Sohta A, Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. 2019. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution* 36: 2069–2085. <https://doi.org/10.1093/molbev/msz131>.
 20. Boussau, Bastien, and Vincent Daubin. 2010. Genomes as documents of evolutionary history. *Trends in Ecology & Evolution* 25: 224–232. <https://doi.org/10.1016/j.tree.2009.09.007>.
 21. Aßmann, Eva, Timo Greiner, Hugues Richard, Matthew Wade, Shelesh Agrawal, Fabian Amman, Sindy Böttcher, et al. 2025. Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. *Nature Water*. <https://doi.org/10.1038/s44221-025-00444-5>.
 22. Gordeev, Victor, Martin Hölzer, Daniel Desirò, Iryna V. Goraichuk, Sergey Knyazev, Helena Solo-Gabriele, Pavel Skums, et al. 2025. Leveraging wastewater sequencing to strengthen global public health surveillance. *BMC Global and Public Health* 3: 23. <https://doi.org/10.1186/s44263-025-00138-w>.
 23. Munteanu, Viorel, Michael Saldana, Nitesh Kumar Sharma, Wenhao O. Ouyang, Eva Aßmann, Victor Gordeev, Nadiia Kasianchuk, et al. 2023. SARS-CoV-2 Wastewater Genomic Surveillance: Approaches, Challenges, and Opportunities. *arXiv.org*. September 23.
 24. Cohen, Ofir, Haim Ashkenazy, Frida Belinky, Dorothée Huchon, and Tal Pupko. 2010. GLOOME: gain loss mapping engine. *Bioinformatics* 26: 2914–2915. <https://doi.org/10.1093/bioinformatics/btq549>.
 25. Nei, M, and T Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3: 418–426. <https://doi.org/10.1093/oxfordjournals.molbev.a040410>.
 26. Arnold O. Allen. 1990. Probability, Statistics, and Queueing Theory. *ScienceDirect*.
 27. Grigelionis, B. 1963. On the Convergence of Sums of Random Step Processes to a Poisson Process. *Theory of Probability & Its Applications* 8. Society for Industrial and Applied Mathematics: 177–182. <https://doi.org/10.1137/1108017>.
 28. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2024. <https://www.r-project.org/>. Accessed August 2.
 29. Sharma, Nitesh Kumar, Ram Ayyala, Dhriti Deshpande, Yesha Patel, Viorel Munteanu, Dumitru Ciorba, Viorel Bostan, et al. 2024. Analytical code sharing practices in biomedical research. *PeerJ Computer Science* 10. PeerJ Inc.: e2066. <https://doi.org/10.7717/peerj-cs.2066>.
 30. Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods* 14. Nature Publishing Group: 1063–1071. <https://doi.org/10.1038/nmeth.4458>.
 31. Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25. Cold Spring Harbor Lab: 1043–1055. <https://doi.org/10.1101/gr.186072.114>.
 32. O’Leary, Nuala A., Matthew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44: D733–D745. <https://doi.org/10.1093/nar/gkv1189>.
 33. Tatusova, Tatiana, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research* 44: 6614–6624. <https://doi.org/10.1093/nar/gkw569>.
 34. Rocha, Jaqueline, Isabel Henriques, Margarita Gomila, and Célia M. Manaia. 2022. Common and distinctive genomic features of *Klebsiella pneumoniae* thriving in the natural environment or in clinical

- settings. *Scientific Reports* 12. Nature Publishing Group: 10441. <https://doi.org/10.1038/s41598-022-14547-6>.
35. Cooper, Helena B., Ben Vezina, Jane Hawkey, Virginie Passet, Sebastián López-Fernández, Jonathan M. Monk, Sylvain Brisse, Kathryn E. Holt, and Kelly L. Wyres. 2024. A validated pangenome-scale metabolic model for the *Klebsiella pneumoniae* species complex. *Microbial Genomics* 10. Microbiology Society,: 001206. <https://doi.org/10.1099/mgen.0.001206>.
 36. Tettelin, Hervé, Vega Massignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome.” *Proceedings of the National Academy of Sciences* 102. Proceedings of the National Academy of Sciences: 13950–13955. <https://doi.org/10.1073/pnas.0506758102>.
 37. Snipen, Lars, and Kristian Hovde Liland. 2015. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* 16: 79. <https://doi.org/10.1186/s12859-015-0517-0>.
 38. Wyres, Kelly L., Margaret M. C. Lam, and Kathryn E. Holt. 2020. Population genomics of *Klebsiella pneumoniae*. *Nature Reviews Microbiology* 18. Nature Publishing Group: 344–359. <https://doi.org/10.1038/s41579-019-0315-1>.
 39. Lam, Margaret M. C., Ryan R. Wick, Stephen C. Watts, Louise T. Cerdeira, Kelly L. Wyres, and Kathryn E. Holt. 2021. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nature Communications* 12. Nature Publishing Group: 4188. <https://doi.org/10.1038/s41467-021-24448-3>.
 40. Akaike, Hirotugu. 1998. A New Look at the Statistical Model Identification. In *Selected Papers of Hirotugu Akaike*, ed. Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, 215–222. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-1694-0_16.
 41. Burnham, Kenneth P., and David R. Anderson, ed. 2004. *Model Selection and Multimodel Inference*. New York, NY: Springer. <https://doi.org/10.1007/b97636>.
 42. Mira, Alex, Howard Ochman, and Nancy A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* 17. Elsevier: 589–596. [https://doi.org/10.1016/S0168-9525\(01\)02447-7](https://doi.org/10.1016/S0168-9525(01)02447-7).
 43. Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405. Nature Publishing Group: 299–304. <https://doi.org/10.1038/35012500>.
 44. Frost, Laura S., Raphael Leplae, Anne O. Summers, and Ariane Toussaint. 2005. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* 3. Nature Publishing Group: 722–732. <https://doi.org/10.1038/nrmicro1235>.
 45. Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16. Nature Publishing Group: 472–482. <https://doi.org/10.1038/nrg3962>.
 46. Ohta, Tomoko. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246. Nature Publishing Group: 96–98. <https://doi.org/10.1038/246096a0>.
 47. Bell, Graham. 2010. Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. Royal Society: 87–97. <https://doi.org/10.1098/rstb.2009.0150>.
 48. Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, et al. 2009. Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics* 5. Public Library of Science: e1000344. <https://doi.org/10.1371/journal.pgen.1000344>.
 49. Rocha, Eduardo P C. 2018. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Molecular Biology and Evolution* 35: 1338–1347. <https://doi.org/10.1093/molbev/msy078>.
 50. Felsenstein, Joseph, and Joseph Felsenstein. 2003. *Inferring Phylogenies*. Oxford, New York: Oxford University Press.
 51. Fitch, Walter M. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 406–416. <https://doi.org/10.2307/2412116>.

52. Huelsenbeck, John P., Rasmus Nielsen, and Jonathan P. Bollback. 2003. Stochastic Mapping of Morphological Characters. *Systematic Biology* 52: 131–158. <https://doi.org/10.1080/10635150390192780>.

LISTA LUCRĂRILOR PUBLICATE LA TEMA TEZEI

Articole în reviste internaționale cotate ISI și SCOPUS:

- 1) LIU, S., RODRIGUEZ, JS., MUNTEANU, V., RONKOWSKI, C., SHARMA, NK., ALSER, M., ANDREACE, F., BLEKHMANN, R., BŁASZCZYK, D., CHIKHI, R., CRANDALL, KA., LIBERA, KD., FRANCIS, D., FROLOVA, A., GANCZ, AS., HUNTLEY, NE., JAISWAL, P., KOSCIOLEK, T., LABAJ, PP., LABAJ, W., LUAN, T., MASON, C., MOUSTAFA, M., MURALIDHARAN, HS., MUTLU, O., GHIASI, NM., RAHNAVARD, A., SUN, F., TIAN, S., TIERNEY, BT., SYOC, EV., VICEDOMINI, R., ZACKULAR JP., ZELIKOVSKY, A., ZELINSKA, K., GANDA, E., DAVERNPORT, ER., POP, M., KOSLICKI, D., MANGUL, S., Analysis of metagenomic data. In: *Nature Reviews Methods Primers*, 2025, vol. 5, pp. 5. ISSN 2662-8449 (Impact Factor 56.0)
- 2) ABMANN, E., GREINER, T., RICHARD, H., WADE, R., AGRAWAL, S., AMMAN, F., BÖTTCHER, S., LACKNER, S., LANDTHALER, M., MANGUL, S., MUNTEANU, V., PSOMOPOULUS, F., SMITH, M., TROFIMOVA, M., ULLRICH, A., VON KLEIST, M., WYLER, E., HÖLZER, M., IRRGANG, G. Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. In: *Nature Water*, 2025, vol. 3, pp. 753-763. ISSN 2731-6084. (Impact Factor 24.1)
- 3) HUANG, YN., JAISWAL, PV., RAJES, A., YADAV, A., YU, D., LIU, F., SCHEG, G., SHIH, E., BOLDIREV, G., NAKASHIDZE, I., SARKAR, A., MEHTA, JH, WANG, K., PATEL, KK., MIRZA, MAB., HAPANI, KC., PENG, Q., AYYALA, R., GUO, R., KAPUR, S., RAMESH, T., CIORBĂ, D., MUNTEANU, V., BOSTAN, V., DIMIAN, M., ABEDALTHAGAFI, MS., MANGUL, S. The systematic assessment of completeness of public metadata accompanying omics studies in the Gene Expression Omnibus data repository. In: *Genome Biology*, 2025, vol. 26, pp. 274, ISSN 1474-760X. (Impact Factor 9.4)
- 4) HUANG, Y., MUNTEANU, V., LOVE, MI., RONKOWSKI, CF., DESHPANDE, D., WONG-BERINGER, A., CORBETT-DETIG, R., DIMIAN, M., MOORE, JH., GARMIRE, LX., REDDY, TBK., BUTTE, AJ., ROBINSON, MD., ESKIN, E., ABEDALTHAGAFI, M., S., MANGUL, S. Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. In: *Cell Genomics*, 2025, vol. 5/5, pp. 100845. ISSN 2666-979X. (Impact Factor 9.0)
- 5) SHARMA, NK., AYYALA, R., DESHPANDE, D., PATEL, Y., MUNTEANU, V., CIORBĂ, D., BOSTAN, V., FICUSTEAN, A., VAHED, M., SAKAR, A., GUO, R., MOOR, A., DARCI-MAHER, N., NOGOY, N., ABEDALTHAGAFI, M., MANGUL, S. Analytical code sharing practices in biomedical research. In: *Peer J Computer Science*, 2024, vol. 10, pp. e20666, ISSN 2376-5992. (Impact Factor 3.8)
- 6) DESHPANDE, D., CHHUNGANI, K., CHANG Y., KARLSBERG, A., LOEFFLER, C., ZHANG, J., MUSZYNSKA, A., MUNTEANU, V., YANG, H., ROTMAN, J., TAO, L., BALLIU, B., TSENG, E., ESKING E., ZHAO, F., MOHAMMADI, P., LABAJ, PP., MANGUL, S. RNA-Seq data science: From raw data to effective interpretation. In: *Frontiers in Genetics*. 2023, vol. 14, pp. 997383. ISSN 1664-8021. (Impact Factor 2.8)
- 7) GORDEEV, V., HÖLZER, M., DESIRÖ, D., GORAICHUK, IV., KNYAZEV, S., SOLOGABRIELE, H., SKUMS, P., KARTHIKEYAN, S., EVANS, A., AGRAWAL, S., LUCACI, AG., MASON, CE., SU, JM., GIBAS, C., NARAJAN, N., PERES DA SILVA, R., DRABCINSKI, N., MUNTEANU, V., ZHAN, L., RUBIN, J., WU, NC., TRISTER, A., CIORBĂ, D., BOSTAN, V., LOBIUC, A., COVASA, M., OPHOFF, RA., ZELIKOVSKY, A., DIMIAN, M., MANGUL, S. Leveraging wastewater sequencing to strengthen global public health surveillance. In: *BMC Global and Public Health*. 2025, vol. 3, pp. 23. ISSN 2731-913X.

Articole în lucrările manifestărilor științifice incluse în bazele de date Web of Science și SCOPUS:

- 1) MUNTEANU, V., LEAHU, A., CIORBĂ, D., CATLABUGA, E., DRABCINSKI, N., DUBCIUC, D., IAPĂSCURTĂ, V., BOSTAN, V. The Pangenome Variability Index: A Quantitative Measure for Assessing Gene Content Diversity in Microbial Genomes. In: *International Conference on Nanotechnologies and Biomedical Engineering*, October 7-10, 2025, Chisinau, Moldova, vol. 2, pp. 1-9, ISBN 978-3-030-31865-9
- 2) BOLDIREV, G., SHARMA, NK., MUNTEANU, V., BHAVATHARINI, A., KOSLICKI, D., ZELIKOVSKY, A., MANGUL, S. Assessing microbial genome representation across various reference databases: A comprehensive evaluation. BioGENext: Next Generation Therapy Conference. September 17-20, 2024, Kyiv, Ukraine. In: *Biopolymers and Cell*. 2024, vol. 40, pp. 169-244, ISSN 1993-6842
- 3) SHARMA, NK., CHHUGANI, K., MUNTEANU, V., SKUMS, P., ZELIKOVSKY, A., MANGUL, S. Realistic assortment of novel metagenomic benchmarks with diverse biological and technological characteristics. BioGENext: Next Generation Therapy Conference. September 17-20, 2024, Kyiv, Ukraine. In: *Biopolymers and Cell*. 2024, vol. 40, pp. 169-244, ISSN 1993-6842

Articole în lucrările manifestărilor științifice incluse în alte baze de date acceptate de ANACEC:

- 1) MUNTEANU, V., DRABCINSKI, N., CIORBĂ, D., BOSTAN, V. Entropy-based Kullback-Leibler Taxonomic Classification of Biological Sequences. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 143-144, ISBN 978-9975-64-480-8
- 2) CATLABUGA, E., DRABCINSKI, N., MUNTEANU, V., SUDACEVSCHI, V. Rare Events Detection and Forecasting in Dynamic Systems. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 167-168, ISBN 978-9975-64-480-8
- 3) MUNTEANU, V., DRABCINSKI, N., CIORBĂ, D., MANGUL, S., BOSTAN, V. The reusability of public omics data across 5 million research publications. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 182-183, ISBN 978-9975-64-480-8
- 4) MUNTEANU, V., CIORBĂ, D., POPIC, V., MANGUL, S. Developing bioinformatics capacity in Moldova. In: *Electronics, Communications and Computing*, October 20-21, 2022, Chisinau, Moldova, pp. 22-23, ISBN 978-9975-45-898-6

Articole în lucrările manifestațiilor științifice incluse în Registrul materialelor publicate în baza manifestărilor științifice organizate din Republica Moldova:

- 1) POPOVA, D., MUNTEANU, V. Large Language Models in Academia: a case study at the Technical University of Moldova. In: *Technical and Scientific Conference for Undergraduate, Master's and Doctoral Students*. Technical University of Moldova, March 27-29, Chisinau, Moldova, vol I, pp. 324-331, ISBN 978-9975-64-458-7
- 2) BAS, A., MUNTEANU, V. Comprehensive assessment of sequence read archive metadata completeness. In: *Technical and Scientific Conference for Undergraduate, Master's and Doctoral Students*. March 27-29, vol II, pp. 1040-1045, ISBN 978 9975-64-460-0

ADNOTARE

la teza cu titlul “**Modele markoviene în timp continuu bazate pe filogenie pentru dinamica genelor în pangenomurile microbiene**”, înaintată de competitorul **MUNTEANU Viorel**, pentru conferirea gradului științific de doctor în informatică, la specialitatea **122.3 “Modelare, metode matematice, produse program”**.

Structura tezei: teza a fost realizată în cadrul Universității Tehnice a Moldovei (UTM), Departamentul Inginerie Software și Automatică, Facultatea Calculatoare, Informatică și Microelectronică. Este scrisă în limba engleză și constă din introducere, 4 capitole, concluzii generale și recomandări, bibliografie din 348 de titluri, 116 text de bază, 47 figuri și 11 tabele. Rezultatele obținute au fost publicate în 16 lucrări științifice, inclusiv: 10 articole recenzate și în reviste cotate ISI și SCOPUS (dintre care 7 cu Factor de Impact); 6 articole în reviste din Registrul Național al revistelor de profil; 3 lucrări prezentate, recenzate și publicate la conferințe naționale și internaționale.

Cuvinte-cheie: bioinformatică, biostatistică, modelare matematică, metagenomică, pangenom, microbiom urban, genomică comparativă.

Scopul lucrării: dezvoltarea unui software bioinformatic reproductibil și scalabil destinat reconstrucției meta-pangenomului din date metagenomice, estimării cantitative a conținutului genic și a ratelor de câștig și pierdere de gene pe arborii filogenetici la nivel de linie taxonomică și clasificării genelor în funcție de presiunea selectivă exercitată de-a lungul acestor linii.

Obiectivele cercetării: (1) dezvoltarea unui software bioinformatic pentru adnotarea genomică, clusterizarea genelor ortoloage, aliniere a secvențelor și inferență filogenetică pentru construcția meta-pangenomului din genomuri asamblate din date metagenomice; (2) definirea și implementarea unui nou model pentru inferența câștigului și pierderii de gene pe arborii filogenetici la nivel de pangenom; (3) dezvoltarea și implementarea unui model statistic pentru detectarea presiunii selecției la nivel de gene și genomuri în pangenomuri; (4) validarea pe seturi empirice, inclusiv metagenom urban și izolate, cu studiu de caz pe genul *Klebsiella* și specia *Klebsiella pneumoniae*.

Noutatea și originalitatea științifică: rezultatele obținute contribuie la soluționarea problemei lipsei unor instrumente computaționale pentru reconstrucția pangenomului și inferența evolutivă a conținutului genomic direct din date metagenomice, prin dezvoltarea unui software bioinformatic scalabil și reproductibil care integrează reconstrucția pangenomului, inferența câștigului și pierderii de gene prin modele Markov cu timp continuu și clasificarea genelor în funcție de presiunea selectivă la nivel de genă și genom, permițând operaționalizarea conceptului de meta-pangenom și analiza integrată a dinamicii fluxului genic și a presiunii selective în comunități microbiene complexe, depășind limitările abordărilor centrate exclusiv pe izolate genomice.

Probleme științifică și de cercetare soluționată: lucrarea introduce un software bioinformatic scalabil și reproductibil pentru reconstrucția pangenomului, care integrează modele Markov în timp continuu și reconstrucția stărilor ancestrale pentru analiza evoluției conținutului genomic direct din date metagenomice. Produsul software dezvoltat permite aplicarea conceptului de meta-pangenom ca extensie a pangenomului, oferind un instrument riguros pentru analiza dinamicii fluxului genic și a presiunii selective asupra genelor în medii complexe.

Semnificația teoretică și valoarea aplicativă a lucrării: lucrarea contribuie la extinderea pangenomului către nivelul de meta-pangenom, oferind o bază metodologică pentru analiza dinamicii fluxului genic și a presiunii selective asupra genelor în medii diverse, inclusiv mediul urban. Din perspectivă aplicativă, software-ul dezvoltat face posibilă utilizarea meta-pangenomului ca instrument de supraveghere genomică, cu relevanță pentru sănătatea publică, agricultură, biotehnologie și abordarea *One Health*, permițând reconstrucția repertoriului genetic din date metagenomice complexe și realizarea de comparații robuste între ecosisteme și clade taxonomice.

Implementarea rezultatelor științifice: Metodele sunt implementate ca set de instrumente bioinformatică cu acces deschis, modulare, scalabile și reproductibile. Acestea sunt utilizat în activități de instruire, implementare în colaborare cu Agenția Națională de Sănătate Publică (ANSP) și laboratoare partenere (Institutul de Microbiologie și Biotehnologie al UTM), prin aplicații pilot de supraveghere genomică urbană și integrarea protocoalelor în fluxuri de lucru operaționale.

ANNOTATION

to the thesis entitled “*Phylogeny based continuous-time markov models for gene dynamics in microbial pangenomes*”, submitted by the candidate **MUNTEANU Viorel** for the award of the scientific degree of Doctor in Computer Sciences, in the specialty **122.3 “Modeling, mathematical methods, software products”**.

Thesis structure: the thesis was carried out at the Technical University of Moldova (TUM), Department of Software Engineering and Automatics, Faculty of Computers, Informatics and Microelectronics. It is written in English and consists of an introduction, 4 chapters, general conclusions and recommendations, a bibliography of 348 sources, 116 pages of main text, 47 figures and 11 tables. The results obtained were published in 16 scientific works, including: 10 peer-reviewed articles in ISI- and SCOPUS-indexed journals (of which 7 with Impact Factor); 6 articles in journals from the National Register of specialized journals; 3 papers presented, peer-reviewed and published at national and international conferences.

Keywords: bioinformatics, biostatistics, mathematical modelling, metagenomics, pangenome, urban microbiome, comparative genomics.

Aim of the work: the development of a reproducible and scalable bioinformatics software intended for the reconstruction of the meta-pangenome from metagenomic data, the estimation of gene counts and gene gain and loss rates on phylogenetic trees at the taxonomic lineage level, and the classification of genes according to the selective pressure acting along these lineages.

Research objectives: (1) the development of a bioinformatics software for genomic annotation, orthologous gene clustering, sequence alignment, and phylogenetic inference for the construction of the meta-pangenome from genomes assembled from metagenomic data; (2) the definition and implementation of a novel model for inferring gene gain and loss on phylogenetic trees at pangenome level; (3) the development and implementation of a statistical model for detecting selective pressure at the gene and genome levels in pangenomes; (4) validation on empirical datasets, including urban metagenomes and isolates, with a case study focusing on the genus *Klebsiella* and the species *Klebsiella pneumoniae*.

Scientific novelty and originality: the obtained results contribute to addressing the lack of computational tools for pangenome reconstruction and evolutionary inference of genomic content directly from metagenomic data, through the development of a scalable and reproducible bioinformatics software that integrates pangenome reconstruction, inference of gene gain and loss using continuous-time Markov models, and gene classification according to selective pressure at the gene and genome levels, thereby enabling the operationalization of the meta-pangenome concept and the integrated analysis of gene flow dynamics and selective pressure in complex microbial communities, overcoming the limitations of approaches centered exclusively on genomic isolates.

Scientific and research problem solved: the thesis introduces a scalable and reproducible bioinformatics software for pangenome reconstruction, which integrates continuous-time Markov models and ancestral state reconstruction to analyze the evolution of genomic content directly from metagenomic data. The developed software product enables the application of the meta-pangenome concept as an extension of the pangenome, providing a rigorous tool for analyzing gene flow dynamics and selective pressure acting on genes in complex environments.

Theoretical significance and practical value of the work: the research contributes to extending the pangenome toward the meta-pangenome level, providing a methodological basis for analyzing gene flow dynamics and selective pressure acting on genes across diverse environments, including urban settings. From an applied perspective, the developed software enables the use of the meta-pangenome as a genomic surveillance tool, with relevance to public health, agriculture, biotechnology, and the One Health approach, allowing reconstruction of genetic repertoires from complex metagenomic data and robust comparisons across ecosystems and taxonomic clades.

Implementation of the scientific results: the methods are implemented as an open-access, modular, scalable, and reproducible set of bioinformatics tools. They are used in training activities and in implementations carried out in collaboration with the National Agency for Public Health (ANSP) and partner laboratories (the Institute of Microbiology and Biotechnology, from TUM), through pilot applications of urban genomic surveillance and the integration of protocols into operational workflows.

АННОТАЦИЯ

к диссертации на тему “**Модели Маркова с непрерывным временем, основанные на филогении, для исследования динамики генов в микробных пангеномах**”, представленной соискателем **МУНТЯНУ Виорелом** на соискание ученой степени доктора по специальности **122.3 “Моделирование, математические методы, программные продукты”** в области компьютерных наук.

Структура диссертации. диссертация выполнена в Техническом Университете Молдовы, на Кафедре Программной Инженерии и Автоматики, Факультете Вычислительной Техники, Информатики и Микроэлектроники. Работа написана на английском языке и включает: введение, 4 глав, общие выводы и рекомендации, библиографию из 348 источников, 116 страниц основного текста, 47 рисунков и 11 таблиц. Полученные результаты опубликованы в 16 научных работах, в том числе: 10 рецензируемых статьях в журналах, индексируемых в ISI и SCOPUS (из них 7 с импакт-фактором); 6 статьях в журналах Национального реестра профильных изданий; 3 докладах, представленных, рецензированных и опубликованных на национальных и международных конференциях.

Ключевые слова: биоинформатика, биостатистика, математическое моделирование, метагеномика, филогенетическая инференция, пангеном, городской микробиом, сравнительная геномика.

Цель работы: разработка воспроизводимого и масштабируемого биоинформатического программного обеспечения, предназначенного для реконструкции, мета-пангенома из метагеномных данных, оценки численности генов и скоростей приобретения и потери генов на филогенетических деревьях на уровне таксономических линий, а также классификации генов в зависимости от действующего вдоль этих линий селективного давления.

Задачи исследования: (1) разработка биоинформатического программного обеспечения для геномной аннотации, кластеризации ортологичных генов, выравнивания последовательностей и филогенетической инференции с целью построения мета-пангенома на основе геномов, собранных из метагеномных данных; (2) разработка и реализация новой модели для инференции процессов приобретения и потери генов на филогенетических деревьях на уровне пангенома; (3) разработка и реализация статистической модели для выявления селективного давления на уровне генов и геномов в пангеномах; (4) валидация на эмпирических наборах данных, включая городские метагеномы и изоляты, с исследованием на примере рода *Klebsiella* и вида *Klebsiella pneumoniae*.

Научная новизна и оригинальность: полученные результаты решают проблему отсутствия вычислительных инструментов для реконструкции пангенома и эволюционной инференции геномного содержания непосредственно из метагеномных данных. В работе разработано масштабируемое и воспроизводимое биоинформатическое программное обеспечение, объединяющее реконструкцию пангенома, инференцию процессов приобретения и потери генов на основе марковских моделей с непрерывным временем, а также классификацию генов по уровню селективного давления, что позволяет проводить интегрированный анализ динамики генетического потока и селективного давления в сложных микробных сообществах.

Решаемая научная проблема: диссертация представляет воспроизводимое биоинформатическое программное обеспечение для реконструкции пангенома, которое интегрирует марковские модели с непрерывным временем и реконструкцию предковых состояний для анализа эволюции геномного содержания непосредственно из метагеномных данных. Разработанный программный продукт обеспечивает применение концепции мета-пангенома как расширения пангенома, предоставляя строгий инструмент для анализа динамики генетического потока и селективного давления.

Теоретическая значимость и практическая ценность: исследование способствует расширению пангенома до уровня мета-пангенома, обеспечивая методологическую основу для анализа динамики генетического потока и селективного давления, действующего на гены в разнообразных средах. Разработанное программное обеспечение позволяет использовать мета-пангеном в качестве инструмента геномного мониторинга, имеющего значение для здравоохранения, сельского хозяйства, биотехнологии и подхода *One Health*, обеспечивая реконструкцию генетических репертуаров из сложных метагеномных данных и проведение устойчивых сравнений между экосистемами и таксономическими кладами.

Внедрение научных результатов: методы реализованы как открытый, модульный, масштабируемый и воспроизводимый набор биоинформатических инструментов. Они используются в учебной деятельности и внедрениях, проводимых в сотрудничестве с Национальным Агентством Общественного Здоровья и партнёрскими лабораториями (Институт Микробиологии и Биотехнологии, ТУМ), через пилотные проекты и интеграцию протоколов в операционные рабочие процессы.

MUNTEANU VIOREL

**MODELE MARKOVIENE ÎN TIMP CONTINUU BAZATE PE
FILOGENIE PENTRU DINAMICA GENELOR ÎN
PANGENOMURILE MICROBIENE**

122.03 Modelare, metode matematice, produse program

Rezumatul științific al tezei de doctor în informatică

Aprobat spre tipar: 25.02.2026
Hârtie offset. Tipar RISO
Coli de tipar 2,5

Formatul hârtiei 60×84 1/16
Tirajul 50 ex.
Comanda nr.

U.T.M., MD-2004, Chișinău, bd. Ștefan cel Mare, 168.
Editura “Tehnica - UTM”
MD-2068, Chișinău, str. Studenților, 9/9