Biomedical Engineering and Technology
Issue 2(22), 2026 ISSN (Online) 2707-8434

УДК 575.113:577.21 DOI: 10.20535/.2026.2(22).363513

МЕТОДИ ПОРІВНЯЛЬНОЇ ГЕНОМІКИ ЯК СУЧАСНИЙ ІНСТРУМЕНТАРІЙ ЕКОЛОГІЧНОЇ ГЕНЕТИКИ

[1]Щербина Валентин Юрійович
shcherbyna-fbmi@lll.kpi.ua
[1]Бертош Наталія Володимирівна
bertosh-fbmi@lll.kpi.ua
[2]Лінник Олена Вячеславівна
elena.linnyk@nure.ua
[1]Галкін Олександр Юрійович
a.galkin@lll.kpi.ua
[1]Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна,
[2]Харківський національний університет радіоелектроніки, м. Харків, Україна.

Анотація. *Порівняльна геноміка є важливим інструментом сучасної екологічної генетики, що дозволяє досліджувати генетичні основи адаптації, еволюції та взаємодії організмів із середовищем. Вона базується на аналізі подібностей і відмінностей геномів різних видів з метою встановлення їх еволюційних і функціональних зв'язків.*

Одними з ключових підходів є методи вирівнювання послідовностей, які дають змогу виявляти гомологічні ділянки ДНК, РНК або білків. Для цього застосовуються як глобальні, так і локальні алгоритми, що забезпечують ефективний аналіз як повних послідовностей, так і окремих їх фрагментів.

Важливим напрямом є ідентифікація ортологічних генів – генів, що походять від спільного предка та зберігають подібні функції. Для цього використовуються філогенетичні, евристичні та комбіновані підходи, а також аналіз синтенії. Кожен із методів має свої переваги та обмеження, що визначає доцільність їх застосування залежно від типу даних і завдань дослідження.

Загалом, методи порівняльної геноміки відіграють важливу роль у розвитку біоінформатики, екологічної генетики та суміжних наук, сприяючи глибшому розумінню закономірностей еволюції та функціонування біологічних систем.

Ключові слова*: порівняльна геноміка, екологічна генетика, екологічна експертиза, методи вирівнювання послідовностей, методи ідентифікації ортологів.*

I. ВСТУП

Екологічна генетика фокусується на встановленні взаємних зв'язків та впливів між генетичними та екологічними процесами. Вперше концепцію екологічної генетики, як генетику популяцій в природних умовах сформував відомий британський вчений Едмунд Бріско Форд, один із співавторів синтетичної теорії еволюції [1]. Методологічно даний міждисциплінарний підхід використовує одночасно широкий арсенал генетичних та екологічних методів. Такого роду дослідження мають важливе значення для розвитку суміжних науково-практичних галузей: еволюційної екології, екологічної експертизи, екологічної токсикології тощо. На сьогодні вивчення екології та еволюції дедалі більше спирається на розуміння геномних основ адаптації, диверсифікації та динаміки популяцій [2]. Геномна структурна мінливість є важливою складовою генетичної мінливості у природних популяціях [3]. Методи порівняльної геноміки знайшли широке застосування у еколого-генетичних дослідженнях [2].

Декодування секвенованих геномів можна порівняти з відновленням інформації з жорсткого диска без знання структури файлів або типу закодованих даних. Водночас геномні дані є значно складнішими, оскільки кінцевий біологічний продукт має багаторівневу організацію та складну функціональну структуру. У цьому контексті порівняльна геноміка виступає потужним інструментом для виявлення та зіставлення особливостей різних геномів. Аналіз подібностей і відмінностей дозволяє

отримувати первинні уявлення про функціональні та еволюційні закономірності, що й становить основу порівняльної геноміки як наукового напряму. [4].

Процеси у порівняльній геноміці часто включають вирівнювання послідовностей ДНК, зокрема використання попарного вирівнювання (алгоритми Нідлмана–Вунша та Сміта–Вотермана) для виявлення тісно споріднених генів. Гомологія є ще одним важливим поняттям, що широко використовується в геномних дослідженнях. Виявлення гомологічних генів має принципове значення у багатьох галузях біології, особливо у порівняльній геноміці. Гомологія означає еволюційний зв’язок між генами, що виникає внаслідок спільного походження, і поділяється на два основні типи залежно від еволюційної події: ортологію (внаслідок видоутворення) та паралогію (внаслідок дуплікації генів).

Ортологічні гени зазвичай ідентифікуються на основі подібності послідовностей; два гени вважаються ортологами, якщо вони є найбільш подібними один до одного серед порівнюваних генів різних видів. Метод взаємних найкращих збігів (reciprocal best hits) та підходи, засновані на аналізі білкових доменів, є стандартними протоколами для виявлення ортологів [5].

На відміну від ортологів, які зазвичай зберігають подібні функції протягом еволюції, паралогічні гени виникають унаслідок дуплікації та з часом можуть набувати нових функціональних властивостей. [6].

II. МЕТА ДОСЛІДЖЕННЯ

Застосування методів порівняльної геноміки постійно розширюється та набуває дедалі більшого значення. У цій роботі проведено аналітичний огляд ключових підходів порівняльної геноміки та наведено їх характеристику.

III. МЕТОДИ ВИРІВНЮВАННЯ ПОСЛІДОВНОСТЕЙ

Вирівнювання послідовностей – біоінформатичний метод, заснований на розміщенні двох або більше послідовностей мономерів ДНК, РНК або білків один під одним таким чином, щоб легко побачити подібні ділянки в цих послідовностях. Подібність первинних структур двох молекул може відображати їх функціональні, структурні або еволюційні взаємозв'язки. Вирівняні послідовності основ нуклеотидів або амінокислот зазвичай подаються у вигляді рядків матриці. Додаються розриви між підставами таким чином, щоб однакові або схожі елементи були розташовані в наступних один за одним шпальтах матриці [7].

Дуже короткі або дуже схожі послідовності можна вирівняти вручну. Однак найбільш цікаві проблеми вимагають вирівнювання довгих, дуже мінливих або надзвичайно численних послідовностей, які неможливо вирівняти виключно зусиллями людини. Натомість людські знання застосовуються при побудові алгоритмів для отримання високоякісних вирівнювання послідовностей, а іноді і при коригуванні кінцевих результатів, щоб відобразити закономірності, які важко представити алгоритмічно (особливо у випадку нуклеотидних послідовностей).

Обчислювальні підходи до вирівнювання послідовностей, як правило, діляться на дві категорії: глобальні вирівнювання та локальні вирівнювання. Розрахунок глобального вирівнювання – це форма глобальної оптимізації, яка "змушує" вирівнювання охоплювати всю довжину всіх послідовностей запитів. На відміну від цього, місцеві вирівнювання ідентифікують регіони подібності в довгих послідовностях, які часто в цілому розходяться. Місцеві вирівнювання часто є кращими, але їх важче обчислити через додаткову проблему визначення регіонів подібності. Різноманітні обчислювальні алгоритми були застосовані до задачі вирівнювання послідовностей. До

них належать повільні, але формально правильні методи, такі як динамічне програмування.

Сюди також належать ефективні евристичні алгоритми або імовірнісні методи, призначені для широкомасштабного пошуку в базі даних, які не гарантують пошуку найкращих відповідностей [8].

Вирівнювання послідовностей зазвичай подають як у графічному, так і в текстовому форматі. У більшості способів представлення послідовності записуються у вигляді рядків, розташованих таким чином, щоб вирівняні залишки знаходилися у відповідних стовпцях. У текстових форматах однакові або подібні символи у вирівняних позиціях можуть позначатися спеціальними символами збереження.

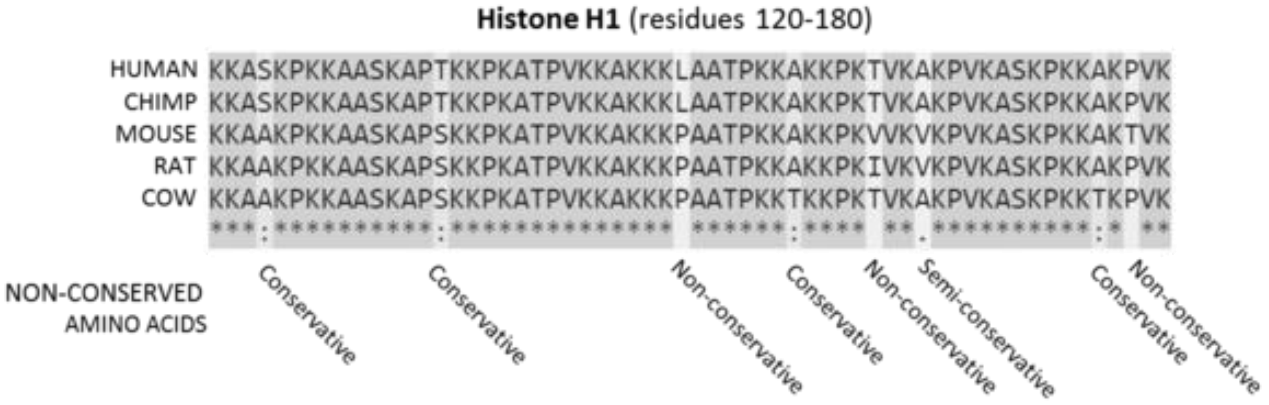

Рис.1 Вирівнювання послідовностей, продукуване ClustalO, білків гістону ссавців.
Залишки, збережені в усіх послідовностях, виділені сірим кольором. Нижче білкових послідовностей наведено ключ, що позначає збережену послідовність (*), консервативні мутації (:), напівконсервативні мутації (.) та неконсервативні мутації () [10].

Для позначення ідентичності між двома стовпцями використовується символ зірочки або труби; інші менш поширені символи включають двокрапку для консервативних підстановок та крапку для напівконсервативних замін. Приклад текстового варіанту зображено на рис.1. Багато програм візуалізації послідовностей також використовують колір для відображення інформації про властивості окремих елементів послідовності; у послідовностях ДНК та РНК це прирівнюється до присвоєння кожному нуклеотиду власного кольору. При вирівнюванні білків, такому як на зображенні вище, колір часто використовується для позначення властивостей амінокислот, що допомагає судити про збереження даної амінокислотної заміни. Для кількох послідовностей останній рядок у кожному стовпці часто є послідовністю консенсусу, що визначається вирівнюванням; послідовність консенсусу також часто представляється у графічному форматі з логотипом послідовності, в якому розмір кожного нуклеотиду або амінокислотної букви відповідає ступеню її збереження. [8, 9].

IV. ПАРНЕ ВИРІВНЮВАННЯ

Парне вирівнювання використовується для знаходження подібних ділянок двох послідовностей. Розрізняють глобальне і локальне вирівнювання. Глобальне вирівнювання передбачає, що послідовності гомологічних по всій довжині. В глобальне вирівнювання включаються обидві вхідні послідовності цілком. Локальне вирівнювання застосовується, якщо послідовності містять як родинні (гомологічні), так і неспоріднені ділянки. Результатом локального вирівнювання є

вибір ділянки в кожної з послідовностей і вирівнювання між цими ділянками [11].

Для отримання парного вирівнювання використовуються різновиди методу динамічного програмування. Зокрема, ці алгоритми реалізовані в сервісах Європейської молекулярно-біологічної лабораторії (EMBL).

Так, наприклад, Needle, Алгоритм глобального вирівнювання, використовує алгоритм Нідлмана - Вунша, а Water, алгоритм локального вирівнювання - алгоритм Сміта - Вотермена.

Порівняння глобального і локального вирівнювань

Для демонстрації в чому відмінність глобального і локального вирівнювань, можна розглянути штучний приклад.

Візьмемо послідовності А і В, і зробимо для них глобальне і локальне вирівнювання. У послідовності була закладена центральна гомологічна ділянка, і помітно відрізняються краї (рис.2).

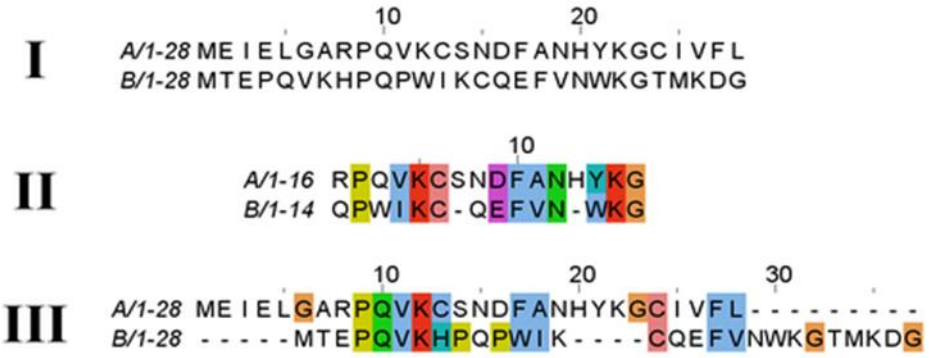

Рис.2 Приклад локального вирівнювання (II; EMBOSS Water.) і глобального (III; EMBOSS Needle.) [10].

Складається алгоритм Нідлмана - Вунша із трьох послідовних етапів [12]:

1. Побудова ініціюючої матриці. Для цього дві порівнювані послідовності розташовують як верхній рядок і як нижні, тобто вони є заголовками матриці. Крім того перед кожною послідовністю виставляють пропуск. І заповнюють перший стовпчик і перший рядок. Заповнення відбувається за допомогою штрафу за пропуски (так як найперше значення в рядку і стовпчику – це пропуск, отже, і перший рядок і стовпчик будуть заповнені від'ємними значеннями).

2. Заповнення таблиці. Заповнення комірки відбувається за такою математичною формулою:

$$F\_ij = \max(F\_i - 1, j - 1 + S(A\_i, B\_j), F\_i, j - 1 + d, F\_i - 1, j + d)$$

де F_{ij} – значення в певній комірці; $S(A_i, B_j)$ – оцінка відповідності між амінокислотами A_iта B_j; d – штраф за пропуск, що задається заздалегідь.

На основі цієї матриці будується матриця локалізації. Слідкують за тим, як відбувалося заповнення, тобто з якої комірки було отримано максимальне значення для наступної комірки.

3. Пошук оптимального вирівнювання.

Процес починається з правої нижньої (кінцевої) комірки матриці та завершується у верхній лівій комірці. Вирівнювання здійснюється шляхом відновлення траєкторії (traceback), яка визначається "вказівками" кожної комірки матриці.

Позначення напрямків мають такий зміст:

- D (diagonal) –перехід по діагоналі, що відповідає вирівнюванню двох символів;
- T (top) – перехід вгору, що біологічно інтерпретується як делеція у горизонтальній послідовності або інсерція у вертикальній;
- L (left) – перехід ліворуч, що відповідає інсерції у горизонтальній або делеції у вертикальній послідовності.

Якщо в комірці матриці зберігається декілька оптимальних значень, це означає наявність кількох можливих напрямків

переходу, тобто альтернативних оптимальних вирівнювань.

V. ЛОКАЛЬНЕ ВИРІВНЮВАННЯ

Використовуються ті ділянки послідовностей, для яких прогнозується максимальна гомологія. Такий підхід є особливо ефективним у випадках, коли лише окремі фрагменти послідовностей є подібними, наприклад унаслідок рекомбінації або конвергентної еволюції. Водночас до коротких ділянок із низьким рівнем подібності слід ставитися обережно, особливо при вирівнюванні довгих послідовностей, оскільки зростає ймовірність випадкового збігу.

За наявності додаткової інформації про подібність функцій пептидів A і B можна припустити, що їхні центральні ділянки є функціонально значущими та можуть визначати загальну функцію відповідних молекул. [13].

Алгоритм Сміта–Вотермана виконує локальне вирівнювання двох послідовностей, враховуючи збіги, невідповідності (заміни), а також вставки та видалення. Вставки й видалення є операціями, що призводять до появи прогалин (gap), які зазвичай позначаються символом "–". Алгоритм Сміта–Вотермана складається з кількох послідовних етапів. [14]:

1. Вибір матриці заміщення та схеми штрафів за розриви. Матриця заміщення визначає оцінки для кожної пари нуклеотидів або амінокислот, відображаючи ступінь їхньої подібності або відмінності. Збіги зазвичай отримують позитивні бали, тоді як невідповідності — нижчі або від'ємні значення. Штрафна система для розривів задає "вартість" відкриття прогалин і їх подовження. Вибір конкретної матриці та параметрів штрафів залежить від мети аналізу, і часто різні комбінації тестуються для отримання оптимального результату.

2. Ініціалізація матриці оцінок.

Формується матриця розміром (1 + довжина першої послідовності) на (1 + довжина другої). Перший рядок і перший стовпець заповнюються нулями. Така ініціалізація дозволяє враховувати локальне вирівнювання без штрафування крайових розривів.

3. Заповнення матриці.

Кожна комірка обчислюється послідовно зліва направо та зверху вниз із врахуванням трьох можливих переходів: діагонального (заміна символів), горизонтального та вертикального (вставки або делеції). Якщо всі отримані значення не є додатними, у клітинку записується 0; інакше обирається максимальне значення, яке також фіксує напрямок переходу.

4. Зворотне відстеження.

Початок вирівнювання визначається з клітинки з максимальним значенням у матриці. Далі шлях відновлюється у зворотному напрямку відповідно до зафіксованих переходів до моменту досягнення нульового значення. У результаті отримують локальні ділянки з максимальною подібністю. Для пошуку наступних варіантів вирівнювання процедуру повторюють, починаючи з інших високих значень поза вже використаними шляхами.

V. МЕТОДИ ІДЕНТИФІКАЦІЇ ОРТОЛОГІВ

Визначення ортологічних генів є ключовим етапом практично всіх порівняльно-геномних досліджень. Такі набори генів застосовуються для аналізу еволюційної консервації та варіабельності молекулярних послідовностей, а також для оцінки швидкостей і механізмів втрати та дуплікації генів. Крім того, вони формують своєрідні "бібліотеки компонентів", які використовуються в системній біології для побудови моделей біологічних систем.

У сучасних порівняльно-геномних підходах мільйони генів з різних секвенованих геномів не розглядаються як незалежні одиниці. Натомість вони групуються у набори передбачуваних ортологів — генів, що походять від спільного

предкового гена та функціонально відповідають “одному й тому самому гену” у різних видів. Такі групи дозволяють реконструювати еволюційну історію генів і переносити функціональні анотації від добре досліджених генів до їхніх менш охарактеризованих гомологів. [15,16].

Проблема ідентифікації ортологічних генів полягає у необхідності відокремлення їх від інших типів гомологічних зв’язків, зокрема паралогії. До основних типів гомологічних зв’язків між генами належать ортологічні та паралогічні, які формуються внаслідок різних еволюційних подій.

Найпоширеніші типи гомологічних взаємозв’язків між генами [17]: Гомологія, Аналогія, Ортологія, Паралогія, Ксенологія, Кооортологія.

Еволюційні події, такі як видоутворення та дуплікація генів, неможливо спостерігати безпосередньо, однак їх можна реконструювати на основі сучасних геномних даних із застосуванням алгоритмічних і статистичних підходів.

Поняття ортології спочатку було введене для опису попарних еволюційних зв’язків між генами, проте на практиці для дослідження еволюції генних сімейств і організмів частіше використовують не окремі пари, а цілі набори ортологів з різних видів. Один і той самий ген може мати різні типи гомологічних зв’язків із різними генами. Наприклад, міоглобін людини є ортологом мишачого міоглобіну, але водночас є паралогічно пов’язаним як із міоглобіном, так і з гемоглобіном миші та людини.

У загальнішому вигляді (див. рис. 3), ген 1α у виду С та ген 1 у виду А є ортологами, оскільки їхнє розходження відбулося внаслідок події видоутворення в спільного предка. Аналогічно, ген 1 у виду А та ген 1β у виду С також є ортологічними. Водночас гени 1α і 1β у межах виду С є паралогами, оскільки виникли в результаті події дуплікації.

Масштабне розрізнення ортологічних і паралогічних генів на основі попередньо визначених наборів передбачуваних ортологів є важливим для реконструкції ключових етапів еволюції та змін у молекулярних функціях. Зокрема, такий підхід дозволив ідентифікувати набір давніх дуплікацій у еукаріотів, для яких характерне збагачення певних функціональних класів генів [15, 18].

VI. ФІЛОГЕНЕТИЧНІ ПІДХОДИ НА ОСНОВІ ДЕРЕВА

Деревовидні методи ґрунтуються на явному представленні еволюційної історії генів у вигляді генеалогічного дерева, що використовується для визначення ортологічних взаємозв’язків. Найбільш прямі підходи передбачають порівняння генетичного дерева з видовим деревом організмів, у яких ці гени виявлені, із застосуванням процедур звірки або картографування дерев [19, 20].

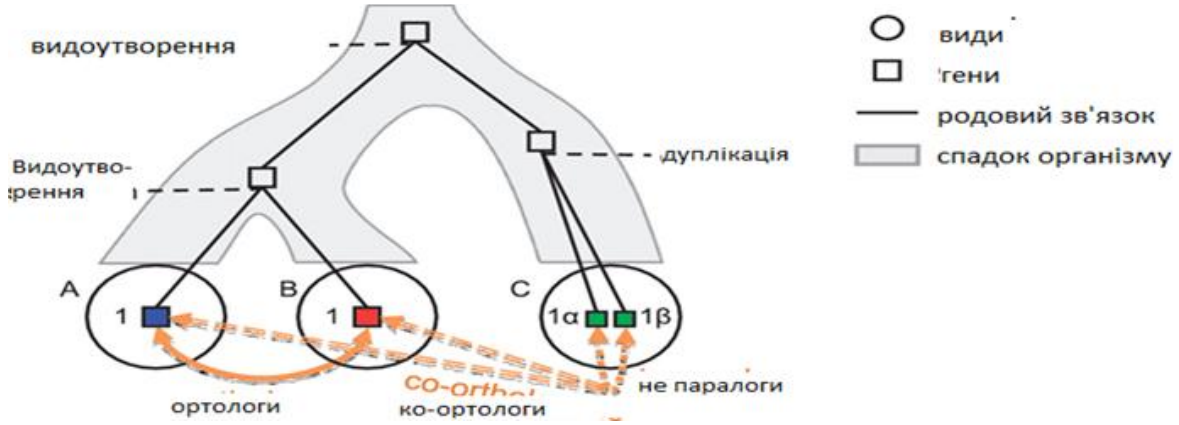

Рис.3 Взаємозв’язки ортології, кооортології та паралогії в еволюції чотирьох генів, що виникли від одного спільного предка [15].

Такий підхід дозволяє зіставити дві еволюційні моделі та ідентифікувати ортологічні гени (рис. 4).Ключовим припущенням цього методу є принцип парсимонії, згідно з яким найімовірнішою вважається еволюційна історія, що передбачає мінімальну кількість подій, таких як дуплікації або втрати генів. Після побудови генетичного дерева класифікація ортологів і паралогів здійснюється на основі їхнього розташування: паралоги зазвичай формують кластери всередині одного виду (рис. 4b), тоді як ортологи групуються з генами інших видів (рис. 4c) [15]. Формально це можна описати так: Якщо нащадки певного вузла в генному дереві розподілені між однаковим набором видів, а наступний вузол охоплює той самий або підмножину цих видів, то між цими вузлами не відбулося події видоутворення, і попередній вузол відповідає дуплікації [15, 19].

Попри те, що філогенетичний аналіз є одним із найточніших підходів для розмежування ортологів і паралогів, він має низку практичних обмежень. Побудова дерев є обчислювально складною при великій кількості генів і організмів, а отримані результати чутливі до шуму та систематичних похибок у даних. Відомими артефактами є ефект довгих гілок і викривлення при аналізі великих або малих еволюційних відстаней. Крім того, точність побудови дерева значною мірою залежить від якості множинного вирівнювання послідовностей, яке може бути ненадійним у випадку багатодоменних білків або великих наборів даних.

Додатковою проблемою є те, що багато методів трактують пропущені ділянки вирівнювання як відсутні дані, що зменшує обсяг інформації для реконструкції еволюційної моделі та може зміщувати інтерпретацію інсерцій і делецій. Перед побудовою дерева також необхідно здійснити відбір гомологічних послідовностей, що саме по собі може вносити додаткові упередження. Використання всіх доступних послідовностей часто є недоцільним через обчислювальну складність і нерівномірність таксономічного представлення. Будь-яка стратегія відбору може спричинити систематичні похибки, які особливо помітні у великих генних сімействах і посилюють труднощі вирівнювання та побудови дерев. У результаті ці обмеження ускладнюють застосування філогенетичних методів до повного набору понад 1000 доступних геномів прокаріотів та еукаріотів [15, 21, 22]

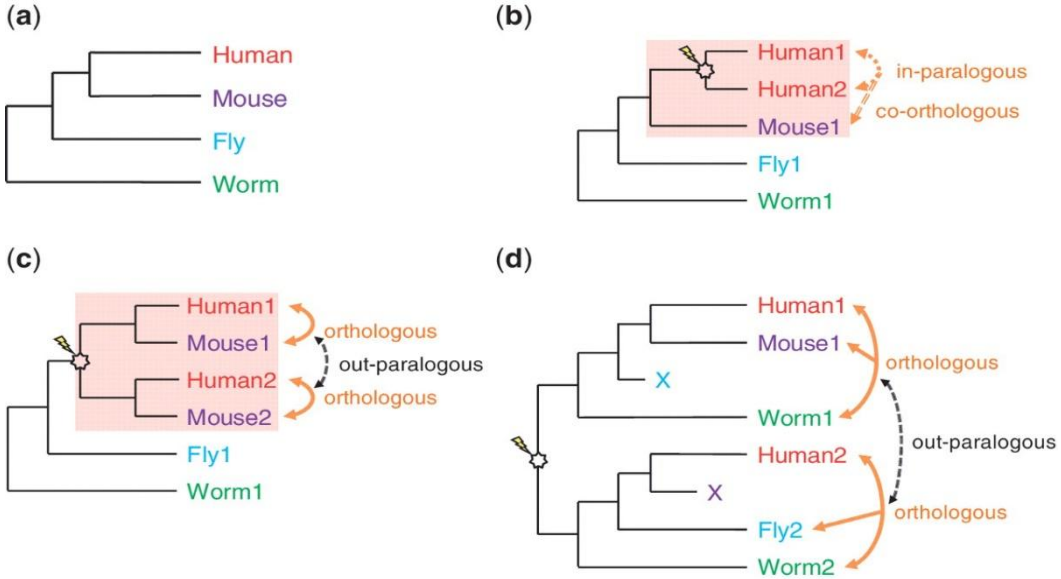

Рис.4 Приклади деревовидних моделей ідентифікації [12, 15]

VII. ЕВРИСТИЧНІ МЕТОДИ НАЙКРАЩОГО ЗБІГУ

На відміну від філогенетичних підходів, які базуються на явному моделюванні еволюції генів і видів, існує інший клас методів, що ґрунтується на припущенні: ортологічні послідовності є більш подібними між собою, ніж до будь-яких інших генів у порівнюваних організмах (рис. 5). У таких підходах основним критерієм є ступінь схожості послідовностей, а не реконструкція еволюційного дерева.

Евристичні алгоритми мають низку переваг порівняно з деревоподібними методами. Вони значно швидші, простіші в автоматизації та добре масштабуються на великі набори геномних даних. Оскільки ці методи не потребують побудови або використання філогенетичних дерев, вони дозволяють уникати помилок, пов'язаних із їх реконструкцією. Крім того, через опору на оцінки подібності, а не на множинні вирівнювання, такі підходи частково обходять проблеми, пов'язані з неточністю вирівнювань і вибором наборів гомологів, що можуть впливати на якість філогенетичного аналізу [15, 23, 24].

Водночас евристичні методи мають власні обмеження. Зокрема, парні порівняння послідовностей часто не дозволяють виявити випадки диференційної втрати генів, які можуть бути коректно інтерпретовані деревоподібними підходами. Також методи, засновані на взаємно найкращих збігах (ВВН), інколи формують надто великі або змішані групи ортологів, що не завжди точно відображає реальну еволюційну історію, особливо у великих і складних генних сімействах.

Додатковим джерелом помилок є рекомбінація доменів. У деяких випадках білки, що містять різні консервативні домени, можуть бути об'єднані в одну групу через наявність проміжних білків із комбінованою структурою (рис. 5c), навіть якщо ці домени не є гомологічними між собою. Оскільки втрати генів і варіабельність багатодоменної архітектури є особливо характерними для еукаріотів, для зменшення таких помилок було запропоновано низку вдосконалених підходів [15, 23, 24, 25].

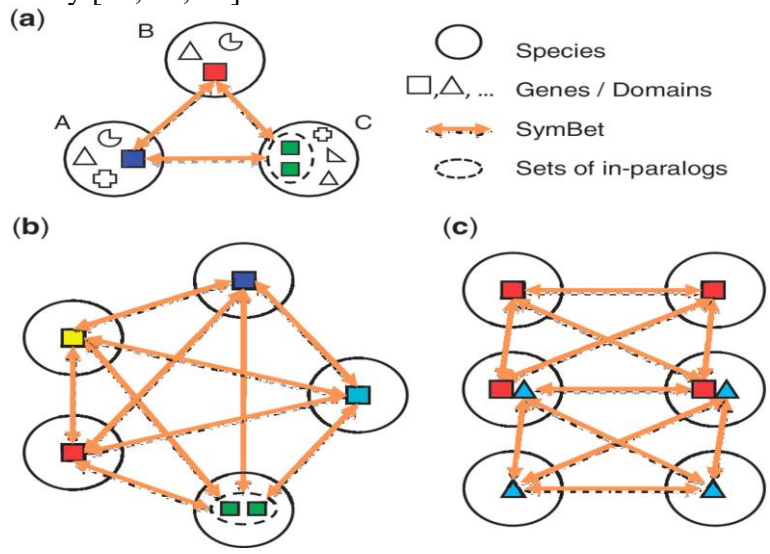

Рис.5 Моделі евристичних методів [12, 15]

VIII. СИНТЕНІЯ

Збереження місцевого генного порядку (синтенії) є наслідком спільного походження, яке найчастіше спостерігається серед близькоспоріднених організмів.

Близько половини всіх ортологічних генів у людини та риб належать до збережених блоків синтетики [19]. У хребетних синтенія виявляється (майже) еволюційно нейтральним за кількома

винятками, хоча темпи геномної перебудови дуже різняться в різних родах. Таким чином, гомологи, оточені наборами ортологічних генів у цих організмах, з великою часткою ймовірності можуть бути самими ортологічними. Однак, принаймні у тварин, швидкість втрати синтетичних сусідств приблизно пропорційна швидкості розбіжності амінокислотних послідовностей в ортологах, і синтезування стає неможливо виявити, коли середня ідентичність білка становить <50% [26]. Прокаріоти демонструють ще вищу швидкість втрати синтезу, яка може відбуватися навіть при> 90% ідентичності [27], за винятком відносно невеликої частки збережених околиць, де тиск селекції діє, щоб зберегти генний порядок [19].

Сама по собі синтенія не є потужним підходом до ідентифікації ортології, оскільки порядки генів, як правило, еволюціонують набагато швидше, ніж генні репертуари або білкові послідовності. Тим не менше, на близьких еволюційних відстанях синтенія може бути використана для підтримки впевненості в прогнозах ортології та навіть допомогти розрізнити ортологію, яка зберігалася вертикально протягом еволюційної історії гена, та ксенологію, отриману в результаті HGT. Інформація про синтенію поєднується з підходом філогенетичного дерева в OrthoParaMap та PhyOP для вимірювання ортології між парою близькоспоріднених видів, а в SYNERGY для використання цієї інформації, коли вона доступна серед великої групи видів. Синтенія також поєднується з підходом до парних зв'язків BBH у невід'ємних щільних геномних кластерах (ATGC) у групах тісно пов'язаних між собою прокаріотичних геномів, а також у MSOAR (згодом розширеному до MultiMSOAR), високопродуктивній системі присвоєння ортологів заснований на перестановці геному, застосованій до ссавців [15, 27, 28].

IX. ГІБРИДНІ ТА ІНШІ ПІДХОДИ

Філогенетичний та евристичний підходи можуть поєднуватися один з одним або з інформацією про синтенію, щоб отримати гібридні підходи, які намагаються подолати недоліки використання будь-якого методу самостійно. Наприклад, гібридні підходи можуть компенсувати обчислювальні витрати філогенетичного підходу або зменшити вразливість евристичних алгоритмів до еволюційних подій, таких як диференціальна втрата генів. Ortholuge використовує філогенетичний підхід для уточнення кластерів, створених за евристичним алгоритмом, відзначаючи випадки, коли відносна дивергенція генів є нетиповою для двох порівняних видів та позагрупових видів, і тому пропонує паралогію, а не ортологію. EnsemblCompara додатково інтегрує підходи до примирення дерев та парних зв'язків BBH, починаючи з генетичних дерев, створених з початкових кластерів, створених евристичними алгоритмами, і узгоджуючи їх з деревом видів хребетних. HomoloGene - ще один гібридний підхід, який використовує попарне порівняння генів, але слідує дереву-керівнику для порівняння більш споріднених організмів, а також додає збереження генетичного сусідства. Існують також інші підходи, які не підпадають під жодну з вищезазначених категорій, включаючи метод, який використовує топологічну відстань у видовому дереві (яке воно не узгоджує з генетичним деревом) як фактор у рівнянні зв’язку для пошуку щільних скупчень у багатосторонній графік (ребра якого не обмежені BBH) та предиктор ортології машинного навчання з використанням набору графічних функцій, які, крім подібності послідовностей та синтетичності, також включають мережі коекспресії генів та взаємодії білків [29,30,31].

X. МЕТОДИ МАШИННОГО НАВЧАННЯ У ПОРІВНЯЛЬНІЙ ГЕНОМІЦІ

Окремим сучасним напрямом розвитку порівняльної геноміки є застосування методів машинного навчання, які дають змогу інтегрувати різнорідні типи геномних і функціональних даних [32]. На відміну від класичних філогенетичних або евристичних підходів, такі методи можуть одночасно враховувати подібність послідовностей, доменну організацію білків, ознаки синтенії, профілі коекспресії генів, мережі білок-білкових взаємодій та інші функціональні характеристики. Це особливо важливо для аналізу складних генних сімейств, багатодоменних білків, випадків дуплікації генів, горизонтального перенесення генетичного матеріалу та неповних або нерівномірно анотованих геномних наборів.
У порівняльній геноміці машинне навчання може використовуватися для прогнозування ортологічних зв'язків, класифікації генних сімейств, виявлення еволюційно консервативних елементів, функціональної анотації генів і пошуку закономірностей у великих пангеномних наборах даних [33-36]. Перспективними є також графові моделі та методи глибинного навчання, які розглядають гени, білки або геноми як елементи складних мереж. Водночас ефективність таких підходів значною мірою залежить від якості навчальних вибірок, повноти анотацій, репрезентативності таксонів і коректного вибору ознак, тому машинне навчання доцільно розглядати не як заміну класичних методів, а як доповнення до філогенетичного, евристичного та синтенійного аналізу.

Узагальнене порівняння підходів до ідентифікації ортологічних генів у Табл 1.

Таблиця 1 Порівняльна характеристика підходів до ідентифікації ортологічних генів

Метод / підхід	Основний принцип	Переваги	Обмеження
Філогенетичні підходи	Побудова генних дерев і їх зіставлення з видовими деревами	Висока специфічність; можливість розмежування ортологів, паралогів, ін-паралогів і аут-паралогів; еволюційно обґрунтована інтерпретація	Висока обчислювальна складність; залежність від якості множинного вирівнювання; чутливість до помилок реконструкції дерев
Евристичні методи найкращого збігу	Визначення ортологів за найбільшою подібністю послідовностей між геномами	Висока швидкість; добра масштабованість; простота автоматизації; придатність для великих геномних наборів	Обмежена здатність враховувати дуплікації, втрати генів і складні еволюційні події; ризик помилок у багатодоменних білках
Аналіз синтенії	Оцінка збереження порядку генів у геномних ділянках	Додаткова підтримка ортологічних прогнозів; корисний для близькоспоріднених організмів; допомагає відрізняти вертикальне успадкування від горизонтального перенесення	Швидка втрата синтенії на великих еволюційних відстанях; обмежена ефективність як самостійного методу
Гібридні підходи	Поєднання філогенетичних, евристичних, синтенійних та інших джерел інформації	Баланс між точністю та масштабованістю; зменшення обмежень окремих методів; можливість інтеграції різних типів даних	Складніша реалізація; залежність від якості кількох джерел даних; потреба у налаштуванні параметрів
Методи машинного навчання	Прогнозування гомологічних або ортологічних зв'язків на основі сукупності ознак	Можливість інтеграції різнорідних даних; перспективність для складних генних сімейств; придатність для великих наборів даних	Залежність від якості навчальних вибірок і анотацій; ризик недостатньої інтерпретованості;

Узагальнення, наведене в таблиці, свідчить, що жоден із підходів не є універсальним для всіх типів порівняльно-геномних задач. Філогенетичні методи забезпечують глибшу еволюційну інтерпретацію, евристичні алгоритми є ефективними для швидкого аналізу великих наборів даних, синтенія підвищує надійність прогнозів у близькоспоріднених організмів, а гібридні та машинно-навчальні підходи дозволяють поєднувати різні джерела інформації. Тому вибір методу має визначатися метою дослідження, таксономічною відстанню між організмами, якістю геномних даних і необхідним рівнем точності.

XI. ВИСНОВОК

Ідентифікація ортологічних генів є одним із ключових завдань порівняльної геноміки, яке ускладнюється нерівномірними темпами еволюції, численними подіями дуплікації та втрати генів, а також горизонтальним перенесенням генетичного матеріалу. Методи визначення пар або груп ортологів умовно поділяються на два основні класи — деревоподібні підходи та евристичні методи, засновані на найкращому збігу; додатково для підвищення точності може використовуватися інформація про геномну синтенію.

Порівняльні дослідження показують, що обидва підходи часто дають схожі результати у вигляді передбачуваних наборів ортологів, а основні розбіжності зазвичай пов'язані з інтерпретацією подій видоутворення, які використовуються для розмежування кооортологів і паралогів. Деревоподібні методи загалом характеризуються вищою специфічністю,

тоді як евристичні підходи забезпечують більшу чутливість.

У теорії перевага надається деревним методам, оскільки вони базуються на явних еволюційних моделях і дозволяють детально класифікувати ортологи, коортологи, ін-паралоги та аут-паралоги. Водночас вони є обчислювально затратними, чутливими до похибок множинного вирівнювання та філогенетичного реконструювання, а також менш ефективними у випадках горизонтального перенесення генів.

Для аналізу великих геномних наборів, особливо у прокаріотів, де еволюційні процеси часто не відповідають простій деревоподібній моделі, більш практичними є швидкі та масштабовані евристичні методи, що ґрунтуються на оцінці подібності послідовностей.

Перспективним напрямом подальшого розвитку порівняльної геноміки є використання методів машинного навчання, які дозволяють інтегрувати послідовнісні, структурні, функціональні та мережеві ознаки для підвищення точності ідентифікації ортологів і функціональної анотації генів. Найбільш доцільним є комбіноване використання таких підходів разом із філогенетичними, евристичними та синтенійними методами.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

Згода на публікацію. Усі автори, які мають відношення до рукопису, дали згоду на публікацію цієї наукової праці.

ORCID ID та внесок авторів.
0009-0007-1594-4812(B, C, D) Valentyn Shcherbyna
0009-0002-9646-0652 (A, B) Nataliia Bertosh
0000-0002-4906-3796 (A, B) Olena Linnyk
0000-0002-5309-6099 (E, F) Alexander Galkin

A – Концепція роботи та дизайн дослідження; B – Аналіз даних; C – Розробка методики дослідження; D – Написання статті; E – Критичний огляд; F – Остаточне схвалення статті

ПЕРЕЛІК ПОСИЛАНЬ

1. Ford E.B. Ecological genetics, 4th ed. Chapman and Hall, London, 1975.
2. Fang B., Edwards S.V. Pangenomes: new tools for ecological and evolutionary genomics // Trends in Ecology & Evolution. – 2026. – Vol. 41, No 3. – P. 230–244. – DOI: 10.1016/j.tree.2025.11.010.
3. Bernatchez L., Ferchaud A.L., Berger C.S., Venney C.J., Xuereb A. Genomics for monitoring and understanding species responses to global climate change // Nature Reviews Genetics. – 2024. – Vol. 25, No 3. – P. 165–183. – DOI: 10.1038/s41576-023-00657-y.
4. Bourque G., Zhang L. Models and Methods in Comparative Genomics // Computational Biology and Bioinformatics. – 2006. – P. 59–104. – DOI: 10.1016/s0065-2458(06)68002-9.
5. Horiike T., Minai R., Miyata D., Nakamura Y., Tateno Y. Ortholog-Finder: A Tool for Constructing an Ortholog Data Set // Genome Biology and Evolution. – 2016. – Vol. 8, No 2. – P. 446–457. – DOI: 10.1093/gbe/evw005.
6. Ong H.S. Comparative Genomics Analysis // Reference Module in Life Sciences. – 2018. – DOI: 10.1016/b978-0-12-809633-8.20126-x.
7. Mount D.M. Bioinformatics: Sequence and Genome Analysis, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2004.
8. Polyanovsky V.O., Roytberg M.A., Tumanyan V.G. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences // Algorithms for Molecular Biology. – 2011. – Vol. 6, No 1. – P. 25. – DOI: 10.1186/1748-7188-6-25.
9. Schneider T.D., Stephens R.M. Sequence logos: a new way to display consensus sequences // Nucleic Acids Research. – 1990. – Vol. 18, No 20. – P. 6097–6100. – DOI: 10.1093/nar/18.20.6097.
10. Sequence alignment [Електронний ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Sequence_alignment
11. Wong K.M., Suchard M.A., Huelsenbeck J.P. Alignment uncertainty and genomic analysis // Science. – 2008. – Vol. 319, No 5862. – P. 473–476. – DOI: 10.1126/science.1151532.
12. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // Journal of Molecular Biology. – 1970. – Vol. 48, No 3. – P. 443–453. – DOI: 10.1016/0022-2836(70)90057-4.
13. Frith M.C. Finding functional sequence elements by multiple local alignment // Nucleic Acids Research. – 2004. – Vol. 32, No 1. – P. 189–200. – DOI: 10.1093/nar/gkh169.
14. Smith T.F., Waterman M.S. Identification of common molecular subsequences // Journal of Molecular Biology. – 1981. – Vol. 147, No 1. – P. 195–197. – DOI: 10.1016/0022-2836(81)90087-5.
15. Kristensen D.M., Wolf Y.I., Mushegian A.R., Koonin E.V. Computational methods for gene orthology inference // Briefings in Bioinformatics. – 2011. – Vol. 12, No 5. – P. 379–391. – DOI: 10.1093/bib/bbr030.
16. Sayers E.W., Barrett T., Benson D.A. et al. Database resources of the National Center for Biotechnology Information // Nucleic Acids Research. – 2011. – Vol. 39. – P. D38–D51.
17. Koonin E.V. Orthologs, paralogs, and evolutionary genomics // Annual Review of Genetics. – 2005. – Vol. 39. – P. 309–338.

18. Makarova K.S., Wolf Y.I., Mekhedov S.L. et al. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell // Nucleic Acids Research. – 2005. – Vol. 33, No 14. – P. 4626–4638.
19. Kawashima T. Comparative and evolutionary genomics // Reference Module in Life Sciences. – 2018. – DOI: 10.1016/b978-0-12-809633-8.20236-7.
20. Page R.D., Charleston M.A. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem // Molecular Phylogenetics and Evolution. – 1997. – Vol. 7, No 2. – P. 231–240.
21. Thompson J.D., Linard B., Lecompte O. et al. A comprehensive benchmark study of multiple sequence alignment methods // PLoS One. – 2011. – Vol. 6, No 3. – e18093.
22. Liu K., Linder C.R., Warnow T. Multiple sequence alignment: a major challenge to large-scale phylogenetics // PLoS Currents. – 2010. – Vol. 2. – RRN1198.
23. Wolf Y.I., Novichkov P.S., Karev G.P. et al. The universal distribution of evolutionary rates of genes // PNAS. – 2009. – Vol. 106, No 18. – P. 7273–7280.
24. Koonin E.V., Wolf Y.I. Genomics of bacteria and archaea // Nucleic Acids Research. – 2008. – Vol. 36, No 21. – P. 6688–6719.
25. Kuzniar A., van Ham R.C., Pongor S. et al. The quest for orthologs // Trends in Genetics. – 2008. – Vol. 24, No 11. – P. 539–551.
26. Zdobnov E.M., Bork P. Quantification of insect genome divergence // Trends in Genetics. – 2007. – Vol. 23, No 1. – P. 16–20.
27. Koonin E.V., Wolf Y.I. Constraints and plasticity in genome evolution // Nature Reviews Genetics. – 2010. – Vol. 11, No 7. – P. 487–498.
28. De Crécy-Lagard V., Hanson A.D. Comparative genomics // Reference Module in Biomedical Sciences. – 2018. – DOI: 10.1016/b978-0-12-801238-3.66095-6.
29. Towfic F., VanderPlas S., Oliver C.A. et al. Detection of gene orthology from gene co-expression // BMC Bioinformatics. – 2010. – Vol. 11 (Suppl. 3). – S7.
30. Vashist A., Kulikowski C.A., Muchnik I. Ortholog clustering on a multipartite graph // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2007. – Vol. 4, No 1. – P. 17–27.
31. Altenhoff A.M., Dessimoz C. Phylogenetic and functional assessment of ortholog inference methods // PLoS Computational Biology. – 2009. – Vol. 5, No 1. – e1000262.
32. Chen Z., Wang L., Zhang Y. et al. From tradition to innovation: conventional and deep learning frameworks in genome annotation // Briefings in Bioinformatics. – 2024. – Vol. 25, No 3. – bbae138. – DOI: 10.1093/bib/bbae138
33. Elnaggar A., Heinzinger M., Dallago C. et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2022. – Vol. 44, No 10. – P. 7112–7127. – DOI: 10.1109/TPAMI.2021.3095381
34. Kirilenko B.M., Munegowda C., Osipova E. et al. Integrating gene annotation with orthology inference at scale // Science. – 2023. – Vol. 380, No 6643. – eabn3107. – DOI: 10.1126/science.abn3107
35. Her H.L., Wu Y.W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains // Bioinformatics. – 2018. – Vol. 34, No 13. – P. i89–i95. – DOI: 10.1093/bioinformatics/bty276
36. Majidian S., Dessimoz C., Gabaldón T. et al. Quest for Orthologs in the era of Data Deluge and AI: Challenges and Innovations in Orthology Prediction and Data Integration // Journal of Molecular Evolution. – 2025. – DOI: 10.1007/s00239-025-10272-6

Biomedical Engineering and Technology
Issue 2(22), 2026 ISSN (Online) 2707-8434

UDC 614.2:004

METHODS OF COMPARATIVE GENOMICS AS A MODERN TOOLKIT OF ECOLOGICAL GENETICS

Valentyn Shcherbyna[1]
shcherbyna-fbmi@lll.kpi.ua
Nataliia Bertosh[1]
bertosh-fbmi@lll.kpi.ua
Olena Linnyk[2]
elena.linnyk@nure.ua
Alexander Galkin[1]
a.galkin@lll.kpi.ua
[1]National Technical University of Ukraine
Ihor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine,
[2]Kharkiv National University Of Radio Electronics
Kharkiv, Ukraine.

Abstract. *Comparative genomics is an important tool of modern ecological genetics, enabling the study of the genetic basis of adaptation, evolution, and interactions between organisms and their environment. It is based on the analysis of similarities and differences between the genomes of different species in order to identify their evolutionary and functional relationships.*
One of the key approaches involves sequence alignment methods, which allow the identification of homologous regions in DNA, RNA, or proteins. Both global and local alignment algorithms are used, providing effective analysis of entire sequences as well as their individual fragments.
An important direction is the identification of orthologous genes—genes that originate from a common ancestor and retain similar functions. For this purpose, phylogenetic, heuristic, and combined approaches are applied, as well as synteny analysis. Each method has its own advantages and limitations, which determine its applicability depending on the type of data and research objectives.
In general, comparative genomics methods play a significant role in the development of bioinformatics, ecological genetics, and related fields, contributing to a deeper understanding of the patterns of evolution and the functioning of biological systems.

Keywords: *comparative genomics, ecological genetics, ecological assessment, sequence alignment methods, ortholog identification methods.*