Біоінформатика та генетичні бази даних

Біоінформатика — це міждисциплінарна галузь науки, що займається розробкою та застосуванням комп'ютерних, математичних і статистичних методів для збирання, зберігання, аналізу та інтерпретації біологічних даних, насамперед генетичної та молекулярно-біологічної інформації.

Основна мета біоінформатики полягає в тому, щоб за допомогою комп'ютерних методів виявляти закономірності у складних біологічних системах, які важко або неможливо дослідити виключно експериментальними методами.

Основні об'єкти дослідження біоінформатики — це ДНК і РНК-послідовності; гени та геноми; білки та їхня структура; метаболічні шляхи; еволюційні зв'язки між організмами; великі масиви даних молекулярної біології (Big Data).

Основні завдання біоінформатики:

- Секвенування та аналіз геномів: визначення послідовності нуклеотидів; пошук генів і їхніх функцій.
- Порівняльна геноміка: порівняння геномів різних організмів; встановлення еволюційної спорідненості.
- Протеоміка — дослідження структури та функцій білків; прогнозування просторової структури білкових молекул.
- Транскриптоміка: аналіз експресії генів; вивчення змін активності генів у різних умовах.
- Біомедична інформатика: пошук генетичних причин захворювань; розробка персоналізованої медицини.
- Методами біоінформатики є вирівнювання послідовностей (BLAST, Clustal Omega); аналіз геномних даних; машинне навчання та штучний інтелект; статистичне моделювання; побудова філогенетичних дерев; молекулярне моделювання.

Біоінформатика в наш час має надзвичайно широке практичне застосування в біології, медицині, сільському господарстві та екології.

1. Медицина та охорона здоров'я

Виявлення генетичних причин спадкових захворювань.

Аналіз мутацій, пов'язаних з онкологічними захворюваннями.

Розробка персоналізованої медицини, коли лікування підбирається на основі геному пацієнта.

Пошук нових лікарських препаратів і вакцин.

Приклад: під час пандемії COVID-19 біоінформатичні методи використовувалися для аналізу геному вірусу SARS-CoV-2 та відстеження появи нових штамів.

2. Генетика та геноміка

Розшифрування геномів організмів.

Пошук генів, що відповідають за певні ознаки.

Аналіз експресії генів у різних тканинах та умовах.

Приклад: визначення генів стійкості до посухи у пшениці.

3. Селекція рослин

Прискорення виведення нових сортів.

Пошук генів, пов’язаних з урожайністю, стійкістю до хвороб і стресових факторів.

Геномна селекція.

Приклад: створення сортів кукурудзи з підвищеною посухостійкістю.

4. Ветеринарія та тваринництво

Виявлення спадкових захворювань у тварин.

Відбір особин із бажаними генетичними характеристиками.

Аналіз збудників інфекційних хвороб.

Приклад: генетичний відбір великої рогатої худоби за продуктивністю молока.

5. Мікробіологія

Ідентифікація бактерій, грибів і вірусів за їх ДНК.

Дослідження мікробіомів ґрунту, води та організму людини.

Вивчення антибіотикорезистентності.

Приклад: аналіз кишкового мікробіому людини для пошуку зв'язку з ожирінням або діабетом.

6. Екологія та охорона природи

Оцінка біорізноманіття за допомогою аналізу ДНК.

Моніторинг рідкісних та зникаючих видів.

Дослідження еволюційних взаємозв'язків.

Приклад: визначення видового складу організмів у водоймі за зразками екологічної ДНК (eDNA).

7. Еволюційна біологія

Побудова філогенетичних дерев.

Визначення спорідненості між видами.

Дослідження походження та еволюції організмів.

Приклад: вивчення еволюційних зв'язків між різними видами дубів (Quercus).

8. Структурна біологія

Передбачення просторової структури білків.

Дослідження взаємодії білків із лікарськими речовинами.

Приклад: використання систем штучного інтелекту для прогнозування структури білків за амінокислотною послідовністю.

Тож, біоінформатика дозволяє перетворювати великі масиви біологічних даних на знання про структуру, функціонування та еволюцію живих систем, що робить її одним із ключових інструментів сучасної біології та біотехнології.

Стрімкий розвиток молекулярної біології, генетики, біоінформатики та високопродуктивного секвенування призвів до накопичення величезних обсягів генетичної інформації. Для її впорядкування, збереження та ефективного аналізу були створені спеціалізовані генетичні бази даних, які сьогодні є невід'ємною складовою сучасної науки та практичної медицини.

Генетична база даних — це структурований набір генетичної інформації, організований таким чином, щоб забезпечувати зручний пошук, аналіз, оновлення та використання даних про нуклеотидні послідовності, гени, білки, геноми, мутації та інші біологічні об'єкти.

Основними завданнями генетичних баз даних є:

- накопичення генетичної інформації;
- стандартизація біологічних даних;
- забезпечення доступу дослідників до результатів наукових робіт;
- проведення порівняльного аналізу геномів;
- підтримка медичних і біотехнологічних досліджень.

Сучасні генетичні бази даних містять інформацію про мільярди нуклеотидних послідовностей та забезпечують глобальний доступ до знань про генетичну організацію живих організмів.

Поява генетичних баз даних безпосередньо пов'язана з розвитком молекулярної біології. Якщо проаналізувати основні етапи розвитку молекулярної біології, то ми побачимо, що однією з перших, ще у 1982 році, виникла база даних GenBank у США. Стрімкий розвиток біоінформатичних технологій та удосконалення методів молекулярно-генетичного аналізу наприкінці 20 століття — на початку 21 століття зумовили накопичення значних обсягів генетичних даних. Починаючи з 2005 коли відбулося впровадження технології NGS (Next Generation Sequencing) відбулося експоненційне зростання кількості генетичної інформації, що одночасно з швидким розвитком хмарних технологій та штучного інтелекту для аналізу даних обумовило появу багатьох різноманітних генетичних баз даних.

Генетичні бази даних можна класифікувати за різними критеріями.

1. Первинні бази даних

Містять експериментально отримані нуклеотидні або білкові послідовності. Приклади: GenBank, EMBL, DDBJ. Їхні особливості — це пряме надходження даних від дослідників; регулярне поповнення та міжнародна стандартизація.

2. Вторинні бази даних

Містять результати аналізу первинних даних. Приклади: InterPro, PFAM, PROSITE. Такі бази містять інформацію про функції білків, консервативні домени та еволюційні взаємозв'язки.

3. Спеціалізовані бази даних

Орієнтовані на певні групи біологічних об'єктів. Наприклад: dbSNP — поліморфізми нуклеотидів; OMIM — спадкові захворювання; ClinVar — клінічно значущі генетичні варіанти; FlyBase — генетика дрозофіли; TAIR — геном Arabidopsis thaliana.

1. GenBank

GenBank є найбільшою відкритою базою нуклеотидних послідовностей. Основні характеристики: підтримується Національним центром біотехнологічної інформації США (NCBI), містить дані всіх основних груп організмів. Вона є безкоштовною та доступною через Інтернет. У кожному записі містяться: назва організму; нуклеотидна послідовність; анотація генів; посилання на наукові публікації.

2. EMBL-ENA

Європейський нуклеотидний архів (European Nucleotide Archive) забезпечує зберігання геномних послідовностей, метагеномних даних, транскриптомів і результатів секвенування.

3. DDBJ

Японська база даних DNA Data Bank of Japan є частиною міжнародного консорціуму INSDC. Вона обмінюється даними з GenBank та EMBL-ENA. Завдяки цьому інформація синхронізується між трьома найбільшими світовими базами даних.

Кожен запис у генетичній базі містить певний набір інформаційних блоків. Основними елементами запису є ідентифікаційний номер та унікальний код запису. Наприклад: NC_000001

Таксономічна інформація містить: вид, рід, родину, порядок, клас, тип (відділ для рослин).

Послідовність представлена у вигляді буквеного коду: ATCGGATCGATCG…

Функціональна анотація описує функцію генів, їхню локалізацію; біологічні процеси та молекулярні функції.

Для накопичення генетичної інформації використовуються: SQL-бази даних, NoSQL-системи, хмарні сховища та спеціалізовані біоінформатичні платформи.

Використовується кілька основних методів аналізу:

- Вирівнювання послідовностей. Дозволяє визначити ступінь подібності між генами різних організмів. Інструменти: BLAST; Clustal Omega; MUSCLE.
- Філогенетичний аналіз. Використовується для дослідження еволюційних зв'язків. Методи: Neighbor Joining; Maximum Likelihood; Bayesian Inference.
- Геномне картування Дозволяє встановити положення генів у геномі.

Генетичні бази даних використовуються практично в усіх сучасних біологічних дослідженнях. Порівняльна геноміка дає змогу порівнювати геноми різних видів, визначати еволюційно консервативні ділянки та встановлювати походження видів.

Дослідження генетичного різноманіття використовуються для аналізу популяцій, міграцій і вивчення генетичної структури популяцій.

Метагеноміка дозволяє аналізувати генетичний матеріал екосистем без культивування організмів. Застосовується для вивчення: ґрунтів, океанів та мікробіому людини.

Медична генетика є однією з головних сфер використання генетичних баз даних. Тут генетичні бази даних застосовуються для діагностики спадкових захворювань. За допомогою баз даних можна визначати патогенні мутації, прогнозувати ризик розвитку хвороб, здійснювати генетичне консультування.

Розвиток персоналізованої медицини також цілком спирається на генетичні бази даних і роботу з ними. Генетичні дані дозволяють обирати оптимальні препарати, визначати дозування та передбачати побічні ефекти. Такий підхід отримав назву фармакогеноміки.

Водночас актуальною є проблема дослідження раку та онкологічних захворювань. Генетичні бази допомагають виявляти мутації BRCA1 і BRCA2 та знаходити генетичні маркери пухлин. Це сприяє ранньому виявленню злоякісних новоутворень.

Сільськогосподарська генетика активно використовує генетичні бази даних для селекції. Геномна селекція дає змогу відбирати рослини на ранніх етапах розвитку, прогнозувати продуктивність і підвищувати врожайність.

Під час створення ГМО саме генетичні бази даних забезпечують пошук необхідних генів, ідентифікацію регуляторних ділянок і контроль трансгенних конструкцій.

Судова генетика використовує ДНК-профілі для ідентифікації особи.

Генетичні бази даних створюють низку етичних проблем. Перша з них – конфіденційність. Генетична інформація є унікальною для кожної людини та потребує особливого захисту. Може виникнути генетична дискримінація. Існують ризики використання генетичної інформації страховими компаніями, роботодавцями і державними структурами.

У найближчі десятиліття очікується інтеграція штучного інтелекту в генетичні бази даних. Це дозволить автоматично аналізувати мутації, прогнозувати функції генів і виявляти генетичні закономірності без участі людини в цих процесах. В найближчому майбутньому очікується створення індивідуальних медичних карт, що міститимуть повні геноми людей. Використання генетичних баз даних стане основою для ранньої діагностики, профілактики захворювань і персоналізованого лікування.

Таким чином, генетичні бази даних є одним із найважливіших інструментів сучасної генетики, біоінформатики та молекулярної біології. Вони забезпечують накопичення, систематизацію й аналіз величезних обсягів генетичної інформації, сприяють розвитку фундаментальних досліджень, медицини, біотехнології та сільського господарства. Завдяки генетичним базам даних стало можливим проводити глобальні геномні дослідження, створювати нові методи діагностики захворювань, розвивати персоналізовану медицину та ефективніше використовувати генетичні ресурси живих організмів. Подальший розвиток інформаційних технологій, штучного інтелекту та геноміки забезпечить ще ширше застосування генетичних баз даних у науковій і практичній діяльності людства.

Дякую

