Лекція 2

Тема. Біоінформатика.

План:

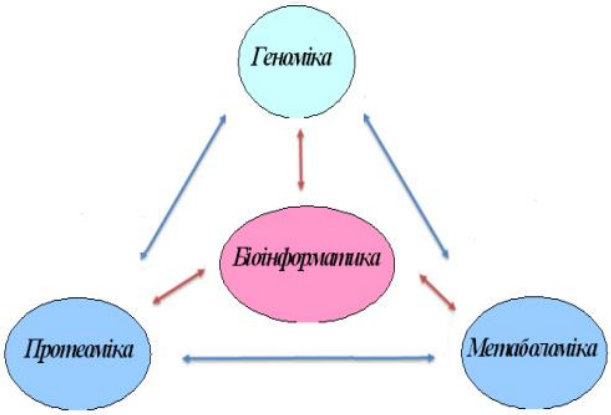
1. Предмет, мета і задачі біоінформатики.
2. Напрямки досліджень біоінформатики
3. Основні галузі досліджень
4. Бази даних

Література: Кеца О. В. Основи біоінформатики: навч.-метод. посібник / О. В. Кеца. – Чернівці : Чернівецький нац. ун-т ім. Ю. Федьковича, 2018. – 192 с.

1. Предмет, мета і задачі біоінформатики.

Біоінформатика – наука, яка вивчає біологічні молекули, використовуючи методи прикладної математики, інформатики та статистики з метою аналізу та впорядкування інформації, пов'язаної з цими біологічними об'єктами. Фактично біоінформатика – це галузь науки, яка здійснює приблизно те саме, що класична біохімія, молекулярна біологія та біотехнологія, однак не в пробірці, а за допомогою обчислювальної техніки.

Засоби біоінформатики широко застосовуються у геноміці, протеоміці, метаболоміці.



Біоінформатика – це шлях від гена до ліків через структуру макромолекули. Все, що реалізувалося раніше за допомогою певних експериментів, зокрема ядерномагнітного резонансу і рентгеноструктурного аналізу, нині можна зробити за допомогою обчислень. Якщо є геном, його можна позначити та знайти кордони гена не за допомогою клонування окремих генів, а за допомогою комп'ютерних програм. Якщо є послідовність білка, можна перейти до просторової структури та функцій, а на основі цих просторових моделей можна сконструювати певні ліки. Біоінформатика набуває застосування у таких напрямах біологічної науки:

– геноміка, транскриптоміка і протеоміка;
– комп'ютерне моделювання в біології розвитку;
– комп'ютерний аналіз генних мереж;
– моделювання в популяційній генетиці.

Біоінформатика – це новітня галузь біології, яка використовує комп'ютерні технології в аналізі та систематизації генетичної інформації для з'ясування структури та функцій макромолекул (ДНК, РНК і білків), складних надмолекулярних комплексів та для моделювання процесів, що відбуваються на молекулярному рівні.

Метою біоінформатики є виявлення структури, функцій та взаємодії біомакромолекул (ДНК, РНК, білків) і подальше використання цих знань при створенні нових лікарський препаратів, наноматеріалів і приладів для діагностики і лікування людини, а також отримання організмів з наперед заданими властивостями.

Задачі біоінформатики:

- аналіз геномів, виділення з їх складу окремих генів, їх екзон-інтронної структури, сигнальних послідовностей,
- передбачення функцій генів і експресованих ними продуктів,
- виявлення генів – потенційних мішеней дії нових ліків,
- оцінка ролі окремих складових амінокислотної послідовності у функціонуванні білка,
- побудова молекулярних моделей білків і нуклеїнових кислот, виходячи з їх послідовностей,
- дослідження механізму функціонування макромолекул, виходячи з їх молекулярних моделей,
- комп'ютерне конструювання ліків, засноване на раціональному виборі генів-мішеней і молекулярних моделей їх білкових продуктів.

2. Напрямки досліджень біоінформатики:

1) *дослідження послідовностей або біоінформатика послідовностей* (послідовності отримують за допомогою традиційних методів **секвенування ДНК:**

- ✓ послідовності ДНК – виявлення кодувальних і некодувальних ділянок послідовності, ідентифікація інтронів та екзонів, передбачення генних продуктів, судово-медична експертиза;
- ✓ геноми – характеристика повторів, структурні особливості генів, філогенетичний аналіз, аналіз генів, відповідальних за певні захворювання;
- ✓ послідовності білків – алгоритми порівняння та множинного вирівнювання послідовностей, ідентифікація консервативних мотивів та функціональних доменів;

2) *дослідження структур або структурна біоінформатика:*

- ✓ досліджуються макромолекулярні структури (білки, РНК, ДНК) – передбачення вторинної і третинної структур, алгоритми вирівнювання 3D-структур, вимірювання геометричних показників білків, міжмолекулярні взаємодії, молекулярні симуляції.

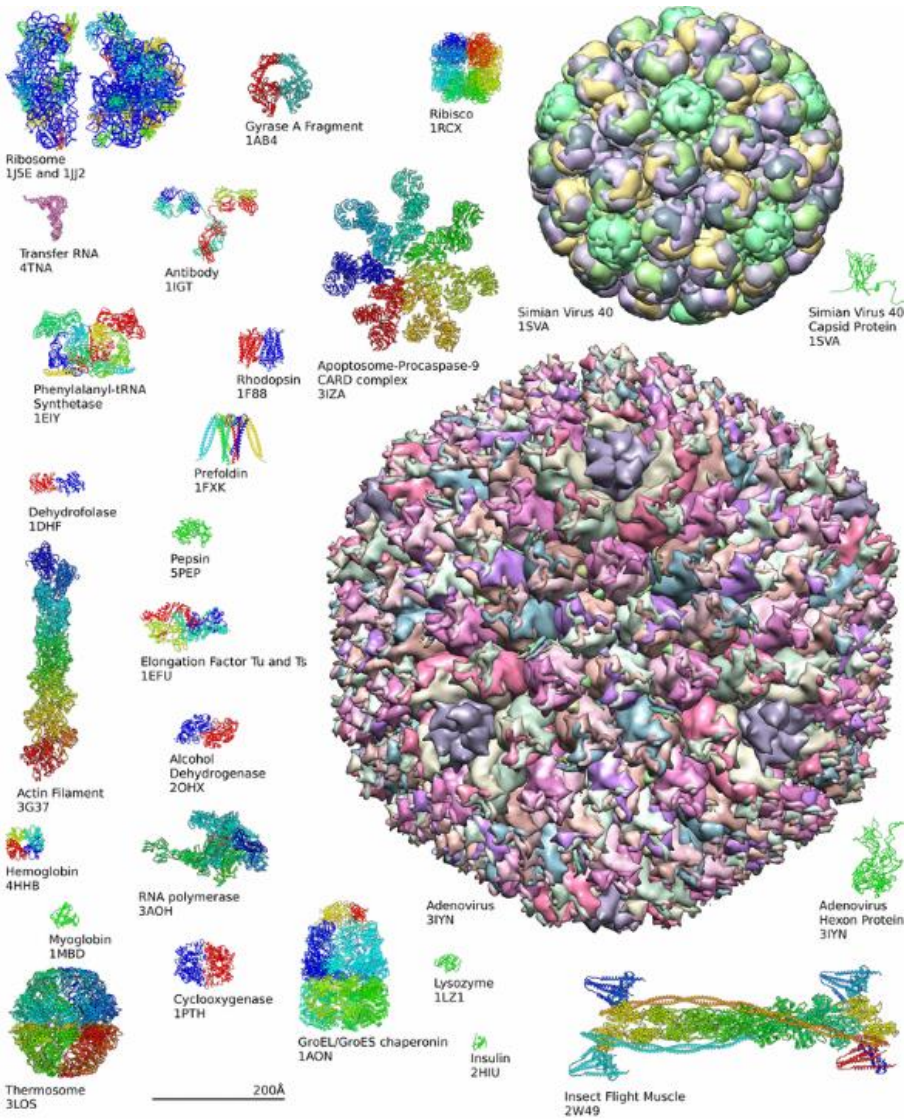
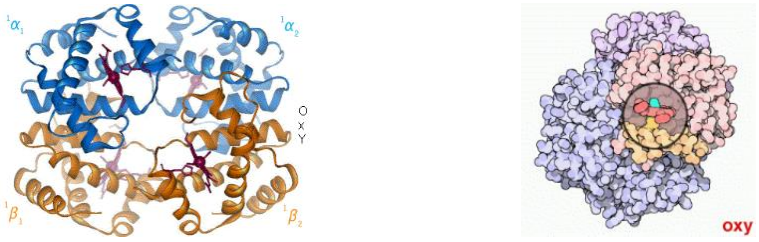


Рис. Приклади структури білків з [Protein Data Bank]



А Б

А) Автор: en:User:BerserkerBen - Uploaded by Habj, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=379538

Б) https://i0.wp.com/www.blopig.com/blog/wp-content/uploads/2016/12/hb-animation.gif?ssl=1

Рис. Просторова структура білків

Структурна біоінформатика використовує методи комп'ютерного аналізу для моделювання просторової структури білків і складних макромолекулярних комплексів (білковобілкових, білково-нуклеїнових) та аналізу механізмів молекулярного розпізнавання.

Метою структурної біоінформатики є також створення селективних модуляторів функціональної активності біополімерів як нових лікарських препаратів;

3) *дослідження метаболітів або метаболічна біоінформатика*:

- ✓ експресія генів – картування даних експресії з даними послідовностей, структури та біохімії;

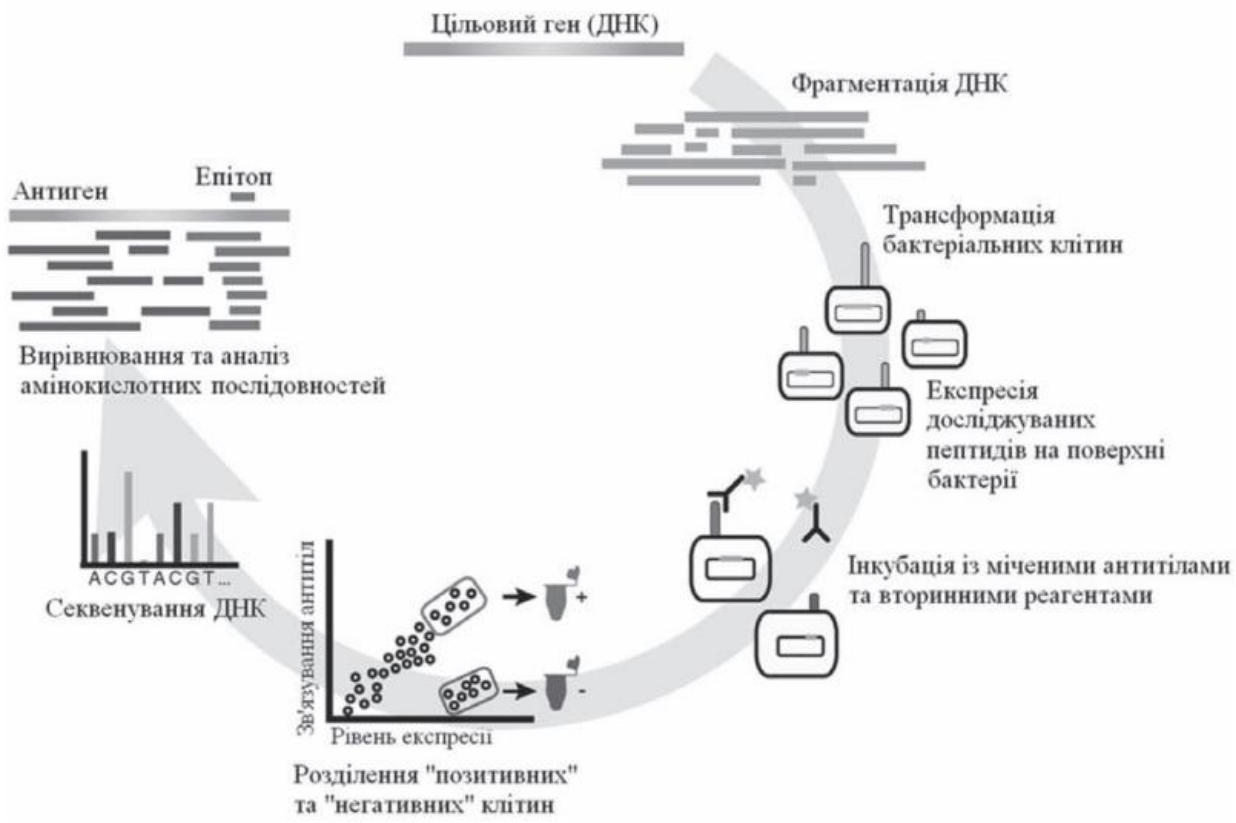


Рис. Принцип епітопного картування із використанням бактеріального дисплею (Галкін О.Ю, doi: http://dx.doi.org/10.15407/ubj86.04.164

- ✓ метаболічні шляхи – симуляція метаболічних шляхів.

У біохімії метаболічний шлях – це зв'язаний ряд хімічних реакцій , що відбуваються в клітині . Реагенти, продукти та проміжні продукти ферментативної реакції відомі як метаболіти, які модифікуються послідовністю хімічних реакцій, що каталізуються ферментами. У більшості випадків метаболічного шляху продукт одного ферменту діє як субстрат для наступного. Однак побічні продукти вважаються відходами та видаляються з клітини. Для функціонування цих ферментів часто потрібні мінерали, вітаміни та інші кофактори.

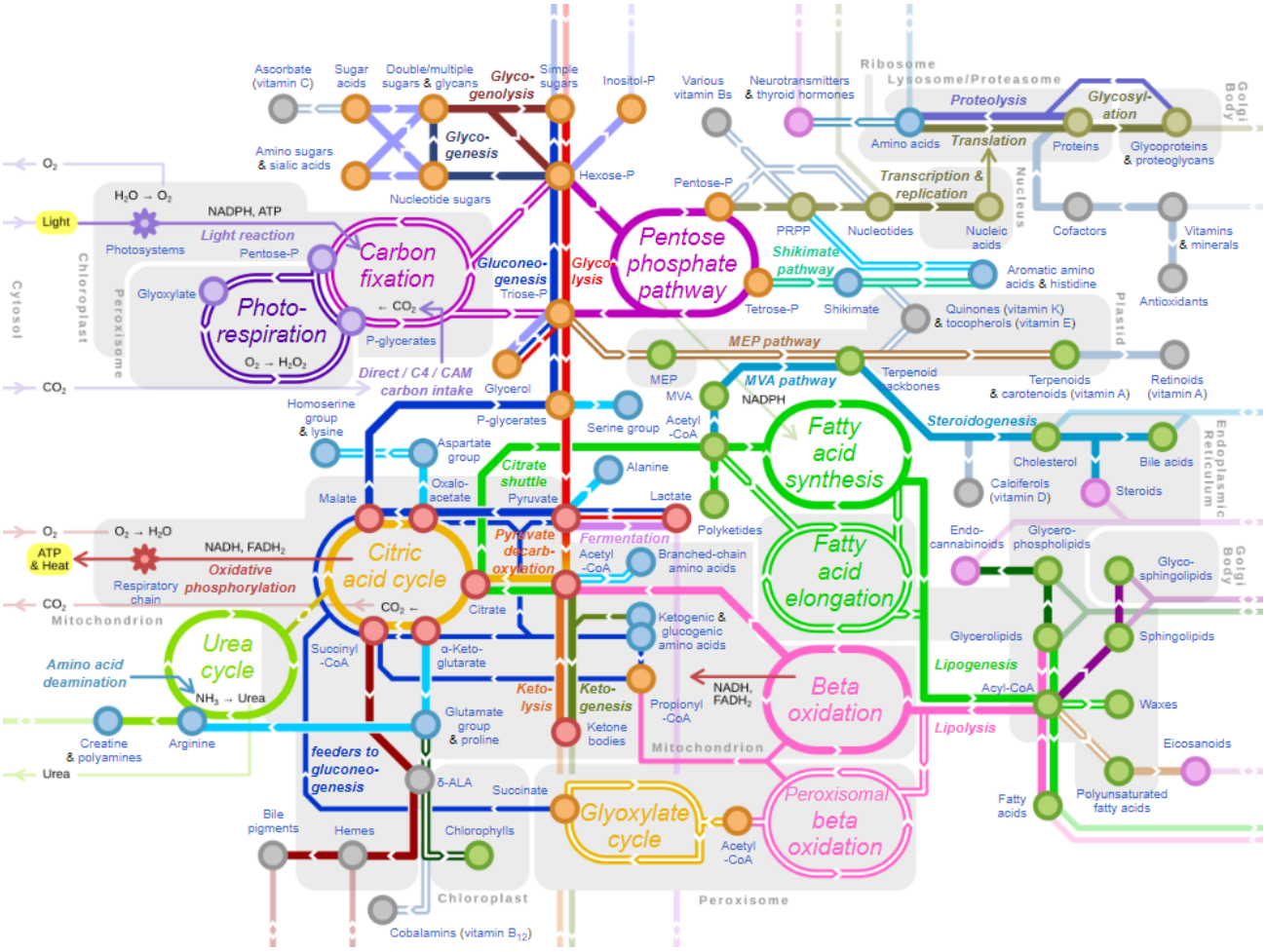
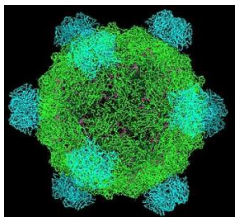


Рис. Метаболічні шляхи

3. Основні галузі досліджень

3.1 Аналіз генетичних послідовностей

У 1977 році був секвенований геном фагу Phi-X174. Ця робота була завершена Фредом Сенгером і його командою в 1977 році. Геном був збережений в базах даних.



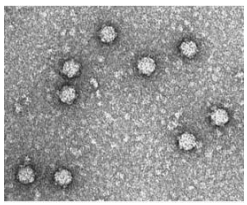


Рис. Бактеріофаг phi X 174 (або ФХ174) — це одноланцюговий ДНК (ssDNA) вірус, який інфікує Escherichia coli (за даними Вікіпедії, https://en.wikipedia.org/wiki/Phi_X_174)

У подальшому розшифровані послідовності ДНК багатьох організмів були збережені в базах даних. Ці дані використовуються для визначення послідовностей білків і регуляторних ділянок. Порівняння генів в рамках одного або різних видів може продемонструвати схожість функцій білків або відношення між видами.

Із зростанням кількості даних вже давно стало неможливим вручну аналізувати послідовності. Для пошуку по геномах тисяч організмів, що складаються з мільярдів пар основ, використовуються комп'ютерні програми. Програми можуть зіставити схожі послідовності ДНК в геномах різних видів, часто такі послідовності несуть схожі функції, а відмінності виникають в результаті дрібних мутацій, таких як заміни окремих нуклеотидів, вставки нуклеотидів і їхнє «випадання» (делецій).

Прикладом секвінування є «дробове секвенування», коли:

1) визначаються послідовності коротких фрагментів ДНК (близько 600–800 нуклеотидів);
2) далі кінці фрагментів накладаються один на одного і, суміщені належним чином, дають повний геном.

У проекті по розшифруванню генома людини збірка зайняла декілька місяців часу суперкомп'ютерів. Зараз цей метод застосовується для практично всіх геномів.

Іншим прикладом застосування комп'ютерного аналізу послідовностей є автоматичний пошук генів і регуляторних послідовностей в геномі. Розробка алгоритмів виявлення ділянок геному, що кодують білки, є важливим завданням сучасної біоінформатики.

Біоінформатика допомагає зв'язати геномні та протеомні проекти, наприклад, допомагаючи у використанні послідовності ДНК для ідентифікації білків.

3.2 Анотація геномів

Анотація – процес маркування генів і інших об'єктів в послідовності ДНК. Перша програмна система анотації геномів була створена в 1995 році Оуеном Вайтом. Доктор Вайт побудував систему знаходження генів, тРНК і інших об'єктів в геномі, і зробив перші позначення функцій цих генів.

3.3 Обчислювальна еволюційна біологія

Еволюційна біологія досліджує походження і появу видів і допомагає еволюційним біологам в декількох аспектах:

1) вивчення еволюції великого числа організмів, вимірюючи зміни в їх ДНК, а не тільки в будові або фізіології
2) порівняння цілих геномів, що дозволяє вивчати більш комплексні еволюційні події, такі як дуплікація генів, горизонтальний перенос генів і передбачати фактори спеціалізації організмів
3) будування комп'ютерних моделей популяцій, щоб передбачити поведінку системи з часом
4) відстеження появи публікацій, що містять інформацію про велику кількість видів

3.4 Оцінка біологічного різноманіття

Спеціалізоване програмне забезпечення застосовується для пошуку, візуалізації і аналізу інформації певного середовища, що складається зі всіх видів, що мешкають в ньому (крапля морської води, жменя землі або вся біосфера планети Земля), і, що важливіше, її доступності іншим людям. Комп'ютерні симуляції моделюють такі речі, як популяційна динаміка, або обчислюють загальне генетичне здоров'я культури в агрономії. Один з найважливіших потенціалів цієї області полягає в аналізі послідовностей ДНК організмів або повних геномів цілих вимираючих видів, дозволяючи запам'ятати результати генетичного

експерименту природи в комп'ютері і можливо використовувати знову в майбутньому, навіть якщо ці види повністю вимруть.

3.5 Аналіз експресії генів

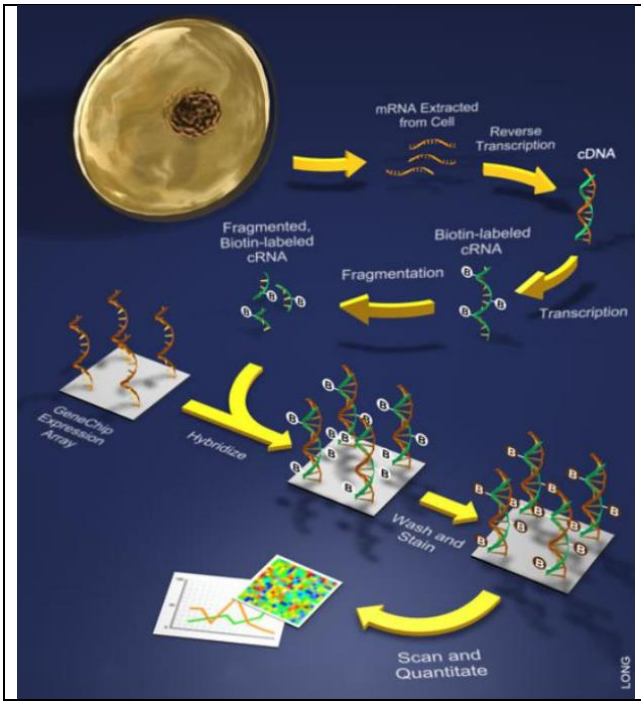
Експресія багатьох генів може досліджуватися за допомогою вимірювання рівнів багатьох мРНК з використанням методів ДНК-мікрочипів, експресії міток послідовностей (EST), серійного аналізу експресії генів (SAGE) або інших варіантів мультиплексної гібридизації in-situ (у біології та біомедичній інженерії in situ означає **дослідження явища саме там, де воно відбувається** (тобто без переміщення його в якесь спеціальне середовище)).

Всі ці методи надзвичайно сприятливі до шуму і схильні до упередженності в отриманих значеннях, тому важлива область досліджень в біофнформатиці займається розробкою статистичних інструментів для розділення сигналу і шуму в генетичних дослідженнях. Ці дослідження часто використовуються для виявлення генів, залучених у хвороби: наприклад, дані мікрочипів ракових епітеліальних клітин порівнюють з нормальними для визначення підвищуючої та понижчуючої регуляції генів.

Метод ДНК-мікрочипів:

ДНК-мікрочіп, або ДНК-чіп - технологія, що використовується в молекулярній біології та медицині. ДНК-мікрочіп це велика кількість невеликих одноланцюгових молекул - ДНК-зондів, які ковалентно пришиті до твердої основи. Кожен такий зонд має суворо певну послідовність нуклеотидів та місце на мікрочіпі. Однакові зонди розташовуються разом, утворюючи сайт мікрочіпа. Між сайтом та послідовністю ДНК зонда є взаємно-однозначна відповідність. ДНК-мікрочіпи використовуються для визначення ДНК або РНК (зазвичай після зворотної транскрипції), які можуть бути білок-кодуючими, так і не кодувати білки. Вимір генної експресії за допомогою кДНК називається профілем експресії, або експресійним аналізом. На сучасних мікрочіпах можна повністю розташувати цілий геном, кожен відомий ген якого буде зондом.

Спосіб вимірювання рівня експресії РНК для кожного гена:



Принцип дії: гібридизація мРНК із зондом – компліментарною ДНК послідовністю закріпленої на твердій поверхні (чіпі).

В основі роботи ДНК-мікрочіпів лежить явище гібридизації. За наявності невеликих кількостей досліджуваного ДНК зразка здійснюють ампліфікацію. Для РНК спочатку

здійснюється зворотна транскрипція, що, втім, необов'язково: існують чіпи, що працюють як з ДНК, так і з РНК. Перевірка зразків ДНК/РНК полягає у міченні зразків різними флуоресцентними мітками для подальшого виявлення та нанесення зразків на мікрочіп. ДНК-мікрочіп з нанесеним на нього зразком інкубують деякий час, щоб відбулася гібридизація комплементарних одноланцюгових молекул, після чого промивають чіп. Усі некомплементарні ДНК/РНК зразка змиваються із чіпа. Після цього сканують мікрочіп за допомогою лазера, який викликає флуоресценцію мічених молекул зразка. Підключений до комп'ютера мікроскоп оцінює флуоресценцію кожного сайту ДНК-мікрочіпа, а отже і встановлює послідовності гібридизованих ДНК, що дозволяє встановити послідовність ДНК, РНК зі зразка

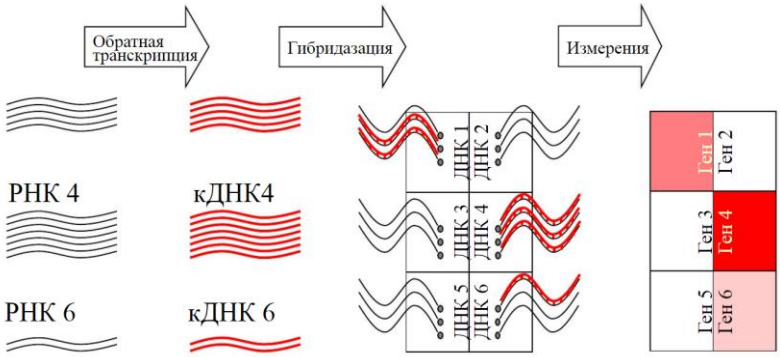


Рис. Робота ДНК-мікрочіпів

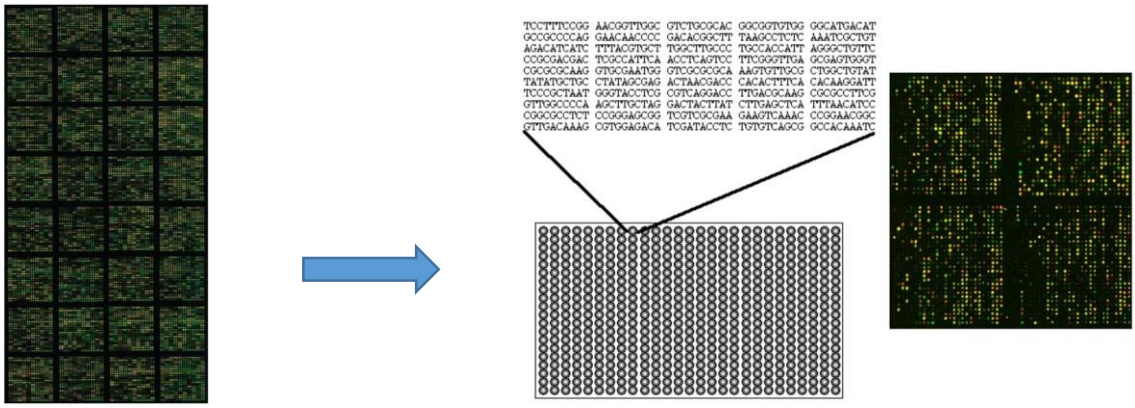


Рис. Двовимірний масив ДНК-зондів для тисяч нуклеотидних послідовностей.

Пояснення:

1) Кожен осередок містить кілька копій певної послідовності ДНК.
2) Можливість оцінки кількості гібридизацій для кожного осередку.
3) Один мікрочіп насправді дозволяє одночасне виконання тисяч експериментів - по одному для кожного гена.
4) Вимірювання експресії генів за різних умов

Рис. Геном миші (праворуч) та людини (ліворуч) на ДНК-мікрочіпах

4. Бази даних

1) Бази даних із біологічною інформацією необхідні для більшості біоінформатичних досліджень. Існує велика кількість таких баз, що містять усе від нуклеотидних послідовностей до опису видів і фенотипів. Багато із них перебувають у вільному доступі, інші закриті. Прикладом вільних баз даних із інформацією про нуклеотидні послідовності є GenBank, DDBJ та ENA (European Nucleotide Archive), сформовані і підтримувані у рамках Міжнародної співпраці баз даних нуклеотидних послідовностей (International Nucleotide Sequence Database Collaboration). Станом на серпень 2014 року GenBank містив 939 775 079 106 пар основ. Інші бази даних більш специфічні, наприклад, присвячені окремому типу генів чи білків (таких як кінази), окремій хромосомі чи органелі або організму. В деяких базах зібрані послідовності об'єднані спільною ознакою, наприклад Pfam (Protein Family) містить кілька тисяч родин гомологічних білків.
2) Бази даних літератури містять бібліографічні дані статей присвячених біологічних дослідженням і посилання на повні тексти статей, одним із найважливіших таких сховищ є MEDLINE.

MEDLINE (**Інтернет-система аналізу та пошуку медичної літератури**, англ. Medical Literature Analysis and Retrieval System Online (MEDLARS Online)) - це бібліографічна база даних про біологічні науки та біомедичну інформацію. Включає бібліографічну інформацію для статей з академічних журналів, що охоплюють медицину, сестринську справу, фармацію, стоматологію, ветеринарію та охорону здоров'я. MEDLINE також охоплює значну частину літератури з біології та біохімії, а також таких галузей, як молекулярна еволюція.

Складений Національною медичною бібліотекою США (НМБ), MEDLINE є у вільному доступі в Інтернеті та доступним для пошуку через PubMed та Національний центр біотехнологічної інформації НМБ, систему Entrez.

MEDLARS - це комп'ютеризована біомедична бібліографічна система пошуку. Вона була запущеною НМБ США в 1964 році. Це було першим масштабним комп'ютерним ретроспективним сервісом пошуку, доступним для широкої громадськості.

База даних містить більше 26 мільйонів записів з 5639 вибраних періодичних видань, що висвітлюють біомедицину та охорону здоров’я з 1950 р до сьогодні. База даних є вільно доступною в Інтернеті через інтерфейс PubMed. У 1995-2003 рр. були додані цитування серед яких 48% припадає на статті, опубліковані в США, близько 88% публікуються англійською мовою, і близько 76% мають тези на англійській мові, написані авторами статей. Найпоширенішою темою в базі даних є рак, із приблизно 12% усіх записів між 1950-2016 рр, які зросли з 6% у 1950 р. до 16% у 2016 р.

Нові журнали не включаються автоматично або негайно. Відбір здійснюється за рекомендаціями комісії, Комітету з технічного огляду відбору літератури, що базується на науковому обсязі та якості журналу.