

Міністерство освіти і науки України  
Чернівецький національний університет  
імені Юрія Федьковича

**О. В. Кеца**

## **ОСНОВИ БІОІНФОРМАТИКИ**

*Навчально-методичний посібник*

Чернівці  
Чернівецький національний університет ім. Ю. Федьковича  
2018

УДК [57.08:004](075.8)

К 377

*Друкується за ухвалою редакційно-видавничої ради  
Чернівецького національного університету  
імені Юрія Федьковича*

К 377 Кеца О. В. Основи біоінформатики: навч.-метод. посібник /  
О. В. Кеца. – Чернівці : Чернівецький нац. ун-т ім. Ю. Федьковича,  
2018. – 192 с.

У навчальному посібнику висвітлено загальні відомості про біоінформаційні засоби аналізу та систематизації біологічної інформації. Окрім теоретичного матеріалу, подані практичні завдання з рекомендаціями, що допоможе студентам самостійно розібратися з основними біоінформаційними ресурсами. Теоретичний та практичний матеріал викладений так, щоб користувач міг самостійно оволодіти основами біоінформатики.

Для студентів вищих навчальних закладів біологічних спеціальностей усіх форм навчання.

УДК [57.08:004](075.8)

У

© Чернівецький національний  
університет ім. Ю. Федьковича, 2018  
© О.В. Кеца, 2018

---

---

## ПЕРЕДМОВА

Навчальний посібник підготовлений згідно з програмою курсу «Біоінформатика» для студентів денної та заочної форм навчання біологічних спеціальностей вищих навчальних закладів і має за мету допомогти студентам опанувати базові засади біоінформатики.

У навчальному посібнику розглянуті основні біоінформаційні бази даних, принципи роботи з біологічними послідовностями (парне та множинне вирівнювання, побудова філогенетичних дерев), біоінформаційні засоби прогнозування структури та функцій білків. Висвітлені основні стратегії комп'ютерного моделювання ліків, передбачених програмою курсу «Біоінформатика». Велика увага приділена онлайн-програмам, доступним для будь-якого користувача через мережу Інтернет та які можуть бути значною допомогою студентам-біологам на початкових етапах вивчення цього курсу.

Для полегшення сприйняття й опрацювання значного обсягу фактичного матеріалу в навчальному посібнику вся інформація поділена на два розділи. Кожний містить теоретичний матеріал та практичні завдання із рекомендаціями. З метою систематизації та узагальнення набутої інформації, наприкінці кожної теми запропоновані тестові завдання.

Навчальний посібник забезпечений великою кількістю ілюстрацій, які дають змогу студентам самостійно розібратися з основними біоінформаційними ресурсами.

Структурованість теоретичного та практичного матеріалу зумовлює поетапну роботу студентів і полегшує засвоєння навчальної програми дисципліни «Біоінформатика».

Навчально-методичний посібник призначений для студентів-біологів усіх форм навчання.

---

---

## РОЗДІЛ 1 ІСТОРІЯ ТА ЗАСОБИ БІОІНФОРМАТИКИ

### ТЕМА 1. ВСТУП ДО БІОІНФОРМАТИКИ

**Мета:** з'ясувати місце біоінформатики серед інших біологічних наук. Ознайомитися з історією розвитку та завданнями біоінформатики як науки, яка розвивається в епоху постгеномних технологій.

#### ТЕОРЕТИЧНА ЧАСТИНА

#### **1. Місце біоінформатики в ланцюгу постгеномних технологій**

Біоінформатика – це новітня галузь біології, яка використовує комп'ютерні технології в аналізі та систематизації генетичної інформації для з'ясування структури та функцій макромолекул і складних надмолекулярних комплексів.

Важливим фактором розвитку біоінформатики стала поява інформації про нуклеотидні та амінокислотні послідовності, внаслідок чого вона набула чітких ознак дедуктивної науки, адже знаючи на сьогодні послідовність геному будь-якого біологічного об'єкта, його можна повністю відтворити клонуванням. Водночас, із накопиченням інформації її обсяги досягнули такого рівня, обробка якого вже не могла бути розумно здійснена без залучення математичних засобів аналізу. Все це об'єктивно сприяло виникненню якісно нового наукового напрямку. Останньою краплею в цій невідворотній низці подій стало розшифрування генома людини.

Завдяки розшифровці структури геномів вірусів, мікроорганізмів, еукаріот і геному людини з'явилися гігантські обсяги інформації щодо послідовностей нуклеїнових кислот і білків. Ця інформація лягла в основу досліджень наступних рівнів її реалізації, таких як регуляція транскрипції й альтернативний сплайсинг генів, регуляція генних кластерів і генних мереж, трансляція у поліпептидні послідовності та транспорт білків у клітині, метаболічні та регуляторні шляхи клітини, білково-білкові і білково-нуклеїнові взаємодії. Як певний завершальний етап щодо цього постає завдання створення комп'ютерних моделей прокаріотичної та еукаріотичної клітин.



---

---

Розшифрування геному людини зумовило виникнення багатьох напрямків принципово нового характеру, які об'єднуються під назвою *постгеномні технології* – технології після розшифрування геному.

До постгеномних технологій належать: *геноміка, транскриптоміка, протеоміка та метаболоміка* – як сукупність технічних і методичних підходів для здійснення високоефективного аналізу живих систем; *біоінформатика* – як засіб забезпечення інформаційно-обчислювальної підтримки досліджень; *нанобіотехнологія* – спосіб практичного впровадження отриманих результатів; *системна біологія* – фундаментальна основа комплексного дослідження молекулярних основ життя.

**Геноміка** вивчає загальні принципи будови геномів та їхню структурно-функціональну організацію. Геноміка виконує секвенування (розшифрування нуклеотидних послідовностей ДНК чи РНК), картування та ідентифікацію функцій генів. Перший повністю розшифрований геном – геном бактеріофага φX174, який складається тільки з 5386 пар нуклеотидів. Значний прогрес у розшифруванні геномів почався в 90-х роках минулого століття завдяки розробкам автоматичних секвенаторів нуклеїнових кислот та необхідного програмного забезпечення і створення спеціалізованих центрів секвенування, таких як Sanger Genome Center. А перші патогенні організми, геноми яких секвеновано, – це бактерія *Haemophilus influenzae* та *Mycoplasma genitalium*. Тому геноміка орієнтована на прикладне застосування геномної інформації та робить значний внесок у розв'язання ключових проблем фармакології та медицини.

Отже, методи секвенування ДНК та біоінформатика допомогли дослідникам аналізувати не окремі короткі послідовності нуклеотидів, як це було колись, а й здійснювати масштабні проекти з прочитання й анотації всієї сукупної ДНК певного виду. Основний недолік геноміки: ми знаємо весь геном, але не знаємо, що реально експресується і працює у клітині.

**Транскриптоміка** вивчає організацію та функціонування транскриптома в певних клітинах або організмі. Транскриптом – сукупність всіх транскриптів (видів РНК), які синтезуються

---

однією клітиною або групою клітин, зокрема мРНК і некодувальні РНК. Поняття «транскриптом» може позначати повний набір транскриптів у даному організмі або специфічний набір транскриптів (молекул РНК), наявних у клітинах певного виду.

На відміну від генома, який, як правило, однаковий для всіх клітин певного організму, транскриптом може сильно змінюватися залежно від умов довкілля. З огляду на те, що поняття «транскриптом» охоплює всі транскрипти даної клітини, він також відображає профіль експресії генів у певний період. Найпоширеніший метод вивчення транскриптома – секвенування РНК і використання ДНК-мікрочипів.

Транскрипти – продукти експресії генів, тому їхня сукупність являє собою перший рівень фенотипу, тобто перший рівень реалізації генетичної інформації, вміщеної в геномі. Структура транскриптому складно організована і постійно змінюється, оскільки залежить від: стадій клітинного циклу, типу клітин і тканин, стадій розвитку організму, стану (норма-хвороба) тканин і органів, наявності зовнішніх сигналів як для самої транскрипції генів, так і для різних посттранскрипційних процесів.

Завдання транскриптоміки – дослідження структури та динаміки транскриптома, що лежить в основі формування другого рівня фенотипу – протеома клітин і тканин. Деякі завдання транскриптоміки водночас є завданнями і функціональної геноміки. Як-от, завдання функціональної геноміки перекриваються із завданнями транскриптоміки та полягають у виявленні та дослідженні умов для формування характерної для клітин певних типів структури транскриптома, зумовленої генетичними програмами розвитку.

Ці умови реалізуються через взаємодію клітинної біохімічної машини з успадкованими разом із геномом регуляторними сигналами: 1) транскрипції; 2) сплайсингу; 3) посттранскрипційних процесів.

Застосування методів біоінформатики на рівні аналізу транскриптома, а саме при дослідженні структури транскриптів і їхнього диференційного часового та просторового розподілу в клітинах і організмах, дає змогу: реконструювати коди,

---

закладені в геномі (кооперація з геномікою); виявляти інформацію у формі сигналів і кодів, необхідну для формування протеома (кооперація з протеомікою).

**Протеоміка** – наука, яка вивчає білковий склад біологічних об'єктів, а також модифікації та структурно-функціональні властивості білкових молекул. Протеомний аналіз спрямований на одночасне вивчення багатьох індивідуальних білків, сукупність яких становить певну систему, яка характеризує досліджуваний об'єкт в цілому. Після розшифрування геномів людини і багатьох інших організмів з'явилися вичерпні бази даних щодо структури всіх білків людини й інших організмів, а також про молекулярну масу їхніх протеолітичних фрагментів, отриманих у стандартних умовах, що допомагає ідентифікувати білки. Розвиток протеоміки зумовлений використанням високотехнологічних методів, що сприяє визначенню кількості певного білка в зразку, дозволяє ідентифікувати білок, його первинну структуру та посттрансляційні модифікації.

У сучасних умовах велика кількість робіт у протеоміці виконується з використанням таких методів, як двовимірний гель-електрофорез у поліакриламіді, високоефективна рідинна хроматографія та мас-спектрометрія.

Дослідження в галузі протеоміки стосуються не тільки всіх білків у будь-якій клітині, але також і всіх варіантів одного білка, його ізоформ і змін, взаємодії між ними, структурного опису білків і їхніх взаємозв'язків у межах однієї клітини, між різними клітинами в одній тканині, в межах міжтканинних взаємозв'язків в цілому багатоклітинному організмі, тобто клітинних комплексів вищого порядку, в яких реалізується весь геном. Протеоміка оснований на досягненнях високого рівня білкової біохімії і здійснює інвентаризацію білків, тобто реально працюючих молекулярних машин у клітині. Сьогодні можна не лише зчитувати послідовності, але й читати й аналізувати всі модифіковані білки: фосфорильовані, глікозильовані та інші. Найголовніше те, що просто в патологічно змінених тканинах легко визначити диспропорцію між білками.

Спільно геноміка та протеоміка виконують розшифрування білків, які кодуються конкретними генами, а також вивчають їхні функції.

---

**Метаболоміка** – «наймолодша» наука, яка аналізує всю сукупність низькомолекулярних речовин (метаболітів) біологічних об'єктів. Метаболічний профіль – молекулярний фенотип живих систем, котрий відображає інформацію, закладену в геномі та реалізовану на транскриптомному і протеомному рівнях. Аналіз метаболічного профілю крові дає змогу врахувати вплив як внутрішніх (ендогенних), так і зовнішніх (екзогенних) чинників на організм та робить його універсальним і перспективним при клінічному використанні.

Метаболом – єдність низькомолекулярних речовин (метаболітів) живої системи, займає особливе положення, оскільки метаболіти є субстратами, інтермедіатами або продуктами більшості біохімічних реакцій, будівельним матеріалом для всіх макромолекул, зокрема складових генома (нуклеотидів), протеома (амінокислот) і транскриптома (нуклеотидів).

Біоматеріал як джерело енергії і будівельного матеріалу потрапляє в живу систему у формі низькомолекулярних речовин. Як низькомолекулярні речовини виводяться з організму також продукти розпаду. Тому, вивчаючи низькомолекулярні речовини метаболоміка всебічно описує молекулярний фенотип біологічного об'єкта з обмеженою мірою деталізації і є невід'ємною частиною досліджень геномів. Метаболічна профілізація дає інформацію про сотні або тисячі метаболітів за один аналіз, що забезпечується застосуванням сучасних високопродуктивних технологій.

За допомогою досліджень метаболітів, задіяних у біохімічних процесах, метаболомний аналіз може використовуватися в клінічній практиці для ефективної діагностики та передбачення ризику захворювань.

Крім того, метаболоміка стосується не лише біомаркерів, тобто метаболітів ендogenousного походження. У токсикологічних і фармакологічних дослідженнях метаболіти поділяють на ендogenousні й екзогенні. Метаболіти чужорідних субстанцій (ліки, отрути) називають ксенометаболітами або ксенобіотиками. У цій сфері досліджень з'являється інший термін – метабономіка.

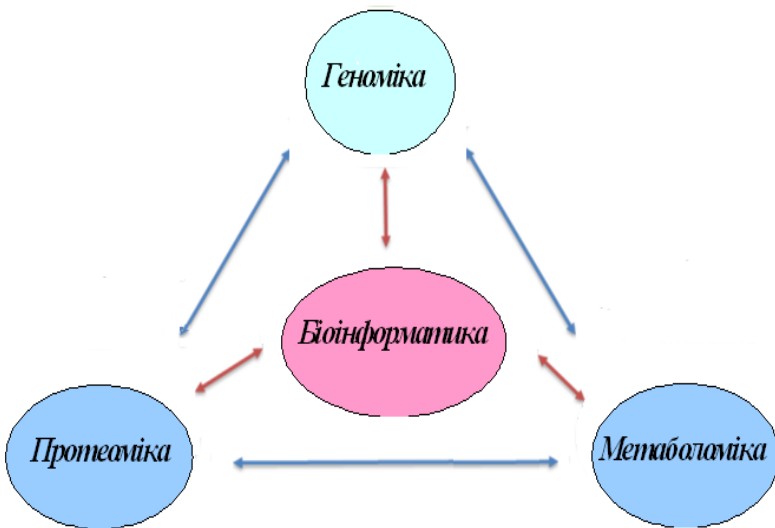
Метаболоміка як самостійна наука найбільшого розвитку набула в тих галузях прикладних досліджень, які, як впливає із самого терміна, спрямовані на визначення різних порушень

---

метаболізму.

Отже, метаболоміка вивчає метаболічні процеси в організмі або конкретній клітині.

Особливу увагу серед постгеномних технологій привертає **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. Це фактично не така галузь, де все вирішує техніка, а інтелектуальна наука, де є місце розуму, думці й освіті. Засоби біоінформатики широко застосовуються у геноміці, протеоміці, метаболоміці.



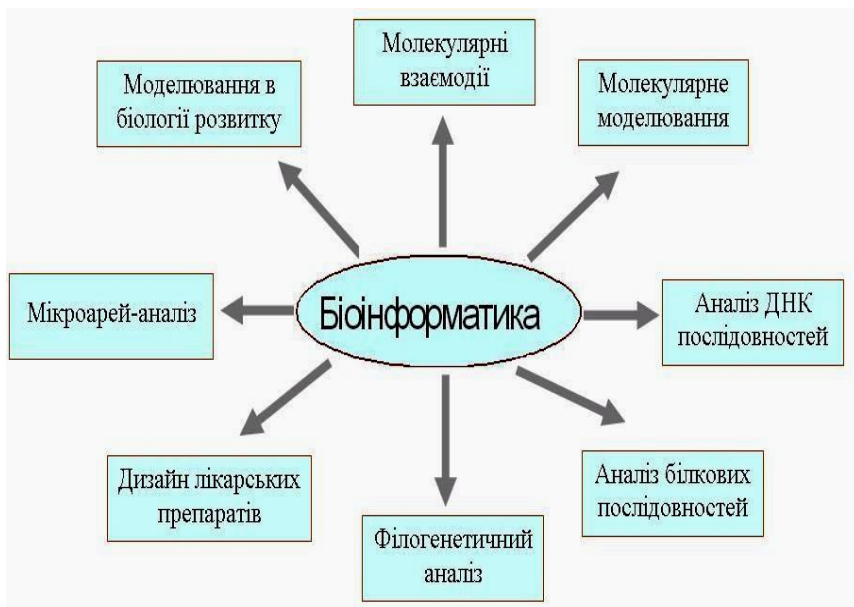
Біоінформатика – це шлях від гена до ліків через структуру макромолекули. Все, що реалізувалося раніше за допомогою певних експериментів, зокрема ядерномагнітного резонансу і рентгеноструктурного аналізу, нині можна зробити за

---

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 напрямках біологічної науки:

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



**Рис. 1. Застосування засобів біоінформатики в різних біологічних науках**

Біоінформатика легко вписалася також у фармакологію, даючи змогу знизити термін проектування препарату з 5–6 років до кількох місяців, а також інтегрувалася в інші медичні та

---

---

біологічні науки.

Нині сформувалися такі розділи біоінформатики: біоінформатика в цілому; клінічна біоінформатика; структурна геноміка; функціональна геноміка; фармакогеноміка; клінічна протеоміка; функціональна протеоміка; структурна протеоміка.

За допомогою методів біоінформатики можна не просто опрацювати величезний масив різних біологічних даних, але і виявляти закономірності, які не завжди встановлюються при звичайному експерименті, передбачати функції генів і зашифрованих в них білків, будувати моделі взаємодії генів у клітині, конструювати лікарські препарати.

Біоінформатика – це новітня галузь біології, яка використовує комп'ютерні технології в аналізі та систематизації генетичної інформації для з'ясування структури та функцій макромолекул та складних надмолекулярних комплексів. Біоінформатика органічно відокремилася з комп'ютерної біології, де комп'ютери та суперкомп'ютери стали самостійним інструментом пізнання в сучасній біології.

## **2. Шляхи розвитку біоінформатики**

Із появи у другій половині ХХ століття, наука інформатика широко впроваджується і співпрацює з іншими науками: фізико-математичними, технічними, гуманітарними.

Наприкінці 60-х – початку 70-х років минулого століття ЕОМ почали активно застосовувати в біології: до того часу вже збільшилася їхня пам'ять і швидкість операцій, зменшилися розміри. Окрім цього, накопичилася велика кількість експериментальних даних із біології, яка вимагала осмислення й обробки.

Сучасна біоінформатика виникла в 70-х роках ХХ століття водночас із появою ефективних методів розшифрування нуклеотидних послідовностей ДНК.

Датою виділення біоінформатики в окрему наукову галузь вважають 1980 рік, коли почалося видавництво журналу *Nucleic Acid Research*, цілком присвяченого комп'ютерним методам аналізу послідовностей.

Важлива віха у становленні та розвитку біоінформатики –

---

проект зі секвенування генома людини. Саме відтоді біоінформатика перестала бути тільки допоміжним інструментом. Перехід до обробки, аналізу та порівняння повних геномів організмів виявився неможливим без використання комп'ютерних методів інформаційного аналізу, отже ці дослідження виділилися у самостійний науковий напрямок.

На початку ХХІ століття відбулася визначна подія у науковому та суспільному житті – прочитана повна нуклеотидна послідовність генома людини, що стало справжнім тріумфом геноміки. На сьогодні відомо кілька сотень повних послідовностей ДНК різних біологічних об'єктів, починаючи з мікоплазми, вірусів, мікроорганізмів і закінчуючи людиною. Поява швидких методів розшифрування послідовності нуклеотидних пар у геномах різних організмів зумовило появу великої кількості генетичних текстів у вигляді довгих символічних послідовностей. Молекула ДНК, носій генетичної інформації, являє собою біополімер, елементарна ланка якого – нуклеотидна пара. Кількість таких пар у ДНК *E. coli* – 5 млн, у дріжджах – 10 млн, у дрозофіли – 100 млн, у людини – близько 3 млрд. Тож найпершим фактором стимулювання розвитку біоінформатики стала необхідність ефективного опрацювання величезних масивів інформації, отриманих експериментальними методами молекулярної біології. А одна з причин створення біоінформатики – підтримка й оновлення баз даних з молекулярної біології, інформація в які надходить із лабораторій усього світу. Починаючи з 80-х років минулого століття створені перші бази даних нуклеотидних послідовностей, вирішені питання зберігання, поповнення та забезпечення доступу до баз даних ДНК- та РНК-послідовностей та їхня обробка. Тобто на цьому етапі біоінформаційні ресурси використовувалися для інвентаризації всіх генів організму людини, які визначають життєдіяльність організму.

Однак для розв'язання задач біології та медицини, а саме створення нових лікарських препаратів, своєчасної діагностики більшості хвороб, добору нових високоселективних ліків, а також для отримання організму з наперед заданими властивостями недостатньо прочитати всі літери генетичного



---

---

коду та визначити необхідний комплект генів, який бере участь у життєдіяльності організму. Тому наприкінці ХХ століття почали інвентаризацію білків – основних функціональних структур організму. Порівняння протеом різних клітин у нормі і при патології дає змогу визначити механізми, які впливають на розвиток патологічних реакцій. Тож другим фактором стимулювання подальшого розвитку біоінформатики став розвиток молекулярної біології, протеоміки, функціональної геноміки, біохімії тощо. Прогрес цих наук став неможливим без сучасних комп'ютерних методів, як-от зіставлення генетичних даних та на основі їхнього аналізу формування теоретичних припущень, які можна перевірити дослідним шляхом та які сприяють отриманню нової інформацію.

Досить часто виникає необхідність спрогнозувати функцію нового білка, що можливо завдяки знаходженню в базах даних схожого, уже вивченого білка. У разі, якщо подібних білків у базах даних немає, здійснюється пошук аналогічних структурних елементів. Тому велику роль у прогресі біоінформатики мають бази даних нуклеотидних і амінокислотних послідовностей не тільки щодо створення, підтримки й оновлення баз даних, а і можливостей зіставляти такі послідовності.

Третім фактором розвитку біоінформатики стало те, що нині передній фронт біологічних і медичних досліджень змістився із мікро- до наномасштабів та нанотехнологій, що знову таки значною мірою пов'язано з розвитком геноміки та протеоміки та сприяло виявленню нових інструментів та методів роботи з найважливішими біологічними макромолекулами, які стосуються наномасштабів – ДНК, РНК та білків.

Перехід до нанобіотехнології, тобто від мікро- до нанорівня, та подальше вдосконалення методів дослідження молекулярно-біологічних даних, розвиток протеоміки, функціональної геноміки, метаболоміки поставив перед біоінформатикою такі задачі:

1. Аналіз геномів, виділення з їхнього складу окремих генів, екзон-інтронної структури та сигнальних послідовностей;
2. Виявлення та передбачення функцій генів і продуктів їхньої експресії як потенційних мішеней для нових ліків;

---

3. Оцінка ролі окремих складових амінокислотної послідовності у функціонуванні білків;

4. Побудова молекулярних моделей білків і нуклеїнових кислот, з урахуванням їхніх послідовностей;

5. Дослідження механізму функціонування макромолекул зважаючи на їхні молекулярні моделі;

6. Комп'ютерне конструювання ліків, основане на раціональному виборі генів-мішеней і молекулярних моделей їхніх білкових продуктів.

Отже, біоінформатика – це галузь науки, яка розробляє і застосовує технології інформатики для: аналізу та систематизації молекулярно-біологічних даних; моделювання процесів, котрі відбуваються на молекулярному рівні з метою виявлення структур, функцій та взаємодії макромолекул із подальшим використанням цих знань у створенні нових лікарських препаратів та нановиробів для діагностики та лікування; отримання організмів із наперед заданими властивостями.

### **3. Мета, завдання та напрямки біоінформатики**

Біоінформатика – наука, яка вивчає біологічні об'єкти, використовуючи методи математики й інформатики для аналізу біологічних процесів.

Основний принцип біоінформатики полягає в тому, що біополімери, наприклад молекули нуклеїнових кислот і білків, можуть мати вигляд послідовності цифрових символів. Окрім того, для відображення мономерів амінокислотних і нуклеотидних ланцюгів необхідна лише обмежена кількість алфавітних знаків.

Основна частина біоінформаційної діяльності сфокусована на аналізі даних, пов'язаних із процесами на молекулярному рівні, та ґрунтується на таких парадигмах:

- послідовність ДНК визначає послідовність амінокислот;
- послідовність амінокислот визначає структуру білка;
- структура білка визначає функцію білка.

Водночас ці парадигми не охоплюють рівні, вищі за молекулярний рівень структурної організації, зокрема, такі

---

---

питання, як тканини стають спеціалізованими протягом розвитку, чи більш загальні – як навколишнє середовище здійснює вплив на генетичні події.

Подібна гнучкість аналізу біомолекул за допомогою обмежених алфавітів зумовила успішне становлення біоінформатики. Розвиток і функціональна потужність біоінформатики значною мірою залежить від прогресу в сфері розробки комп'ютерних апаратних засобів і програмного забезпечення.

Найпростіші завдання біоінформатики стосуються створення та ведення баз даних біологічної інформації.

*Предмет* біоінформатики складається із трьох компонентів:

- 1) створення баз даних, які дають змогу здійснювати зберігання великих наборів біологічних даних і керування ними;
- 2) розробка алгоритмів і методів статистичного аналізу для визначення відношення між елементами баз даних;
- 3) використання цих засобів для аналізу й інтерпретації біологічних даних різного типу – зокрема послідовностей ДНК, РНК і білків, білкових структур, профілей експресії генів і метаболічних шляхів.

*Мета* біоінформатики триєдина:

1. Упорядкувати наявні дані так, щоб вони були доступними для інших дослідників (наприклад Protein Data Bank із тривимірними макромолекулярними структурами білків), які могли б вносити в них нову інформацію в міру надходження нових відомостей;

2. Розвивати програмні засоби та інформаційні ресурси, які допомагають аналізувати наявні дані (наприклад, маючи встановлену амінокислотну послідовність білка, можна охарактеризувати її, порівнюючи із уже відомими послідовностями; для цього використовують програми FASTA та PSI-BLAST);

3. Застосовувати засоби та ресурси для аналізу інформації й інтерпретувати результати аналізу та висновки у необхідному для біології вигляді (у біоінформатиці можемо здійснити глобальний аналіз усіх наявних даних з метою виявлення споріднених ознак і унікальних властивостей у багатьох системах).

*Завдання* біоінформатики полягають в аналізі інформації, закодованої в біологічних послідовностях, зокрема:

- виявляти гени в послідовностях ДНК різних організмів;

---

- розвивати методи вивчення структури та функцій нових розшифрованих послідовностей і відповідних структурних ділянок РНК;

- визначати родини споріднених послідовностей і будувати моделі;

- вирівнювати подібні послідовності і будувати філогенетичні дерева для встановлення еволюційних зв'язків.

Окрім вищевказаних завдань, – є ще одне важливе питання біоінформатики, прямо пов'язане з фармакологічною біотехнологією – виявлення мішеней для медикаментозної дії лікарських препаратів і пошук перспективних дослідних засобів.

Предмет біоінформатики реалізується в таких *видах діяльності*:

1. Керування біологічними даними та їхня обробка; сюди належить організація, відстежування, захист, аналіз;

2. Організація зв'язку між ученими, організаціями та проектами, залученими до фундаментальних і прикладних біологічних досліджень. Зв'язок може вмещувати електронну пошту, дистанційний вхід у систему, мережеві інформаційні ресурси;

3. Організація наборів біологічної інформації, документів і літератури, а також забезпечення доступу до них;

4. Аналіз та інтерпретація біологічних даних із використанням обчислювальних методів: візуалізація, математичне моделювання, побудова алгоритмів обробки біологічних структур.

Нині розрізняють кілька основних напрямків біоінформатики:

***Залежно від досліджуваних об'єктів:***

біоінформатика послідовностей:

- *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;

структурна біоінформатика:

- *макромолекулярні структури* – передбачення вторинної і третинної структур, алгоритми вирівнювання 3D-структур, вимірювання геометричних показників білків, міжмолекулярні взаємодії, молекулярні симуляції.

Структурна біоінформатика використовує методи комп'ютерного аналізу для моделювання просторової структури білків і складних макромолекулярних комплексів (білково-білкових, білково-нуклеїнових) та аналізу механізмів молекулярного розпізнавання. Метою структурної біоінформатики є також створення селективних модуляторів функціональної активності біополімерів як нових лікарських препаратів;

метаболична біоінформатика:

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

- *метаболичні шляхи* – симуляція метаболичних шляхів.

Отже, біоінформатика належить до високих технологій сучасної біології, що забезпечує інформаційно-комп'ютерні та теоретичні основи генетики, селекції, молекулярної біології, біохімії, білкової інженерії, біотехнології, медичної генетики, генної діагностики та екології.

---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

**1. Яке з поданих визначень найповніше описує біоінформатику?**

а) теоретична дисципліна, що вивчає структуру, властивості інформації та методи її створення, зберігання та використання;

б) сукупність методів математичного опрацювання даних, одержаних при вимірюванні окремих органів організмів;

в) наука, що вивчає біологічні об'єкти на основі методів прикладної математики, інформатики та статистики для аналізу та впорядкування інформації, пов'язаної з цими об'єктами.

**2. Доповніть речення: Залежно від досліджуваних об'єктів розрізняють такі основні напрями біоінформатики: а) ...; б) ...; в) ...**

**3. Які із зазначених наук належать до постгеномних?**

а) геноміка;

б) протеоміка;

в) біоінформатика;

г) систематика;

д) кібернетика;

е) системна біологія;

є) молекулярна біологія.

**4. Який рік можна вважати датою виділення біоінформатики в окрему наукову галузь?**

а) 1960;

б) 1980;

в) 2001;

г) 2009.

**5. Дайте відповідь, де зберігаються біоінформаційні дані.**

**6. Моделювання просторової структури білків і складних макромолекулярних комплексів вивчає .....**

а) біоінформатика послідовностей;

б) структурна біоінформатика;

в) метаболічна біоінформатика.

**7. Яка наука вивчає загальні принципи будови геномів та їхню структурно-функціональну організацію?**

а) протеоміка;

б) геноміка;

в) метаболоміка;

- 
- 
- г) транскриптоміка;
  - д) біоінформатика.

**8. Бурхливий розвиток біоінформатики пов'язаний із розширенням геному:**

- а) людини;
- б) бактеріофага φX174;
- в) *E. coli*.

**9. Який напрям біоінформатики вивчає структурні особливості генів, філогенетичний аналіз, аналіз генів, відповідальних за певні захворювання?**

- а) біоінформатика послідовностей;
- б) структурна біоінформатика;
- в) метаболічна біоінформатика.

**10. Заповніть пропуски: ..... вивчає організацію та функціонування транскриптома в певних клітинах або організмах.**

**11. Доповніть речення: Сукупність усіх видів РНК, які синтезуються однією клітиною або групою клітин називається ...**

- а) геномом;
- б) транскриптомом;
- в) протеомом.

**12. Чи можна за допомогою біоінформаційних ресурсів будувати філогенетичні дерева з метою виявлення еволюційних зв'язків між організмами?**

- а) так; б) ні.

**13. Картування даних експресії генів із даними послідовностей, структури та біохімії здійснює .....**

- а) біоінформатика послідовностей;
- б) структурна біоінформатика;
- в) метаболічна біоінформатика.

**14. Наука, яка вивчає білковий склад біологічних об'єктів, модифікації і структурно-функціональні властивості білків, називається .....**

- а) геномікою;
- б) транскриптомікою;
- в) метаболомікою;
- г) протеомікою;
- д) біоінформатикою.

---

---

**15. Чи можна вважати розвиток нанотехнологій одним із факторів розвитку біоінформатики?**

а) так; б) ні.

**16. Доповніть відповідь: *Нині велика кількість експериментів в протеоміці виконується з використанням таких методів:*** а) .....; б) .....; в) .....

**17. Наука, яка аналізує всю сукупність низькомолекулярних речовин (метаболітів) біологічних об'єктів називається .....**

- а) геномікою;
- б) транскриптомікою;
- в) протеомікою;
- г) метаболомікою;
- д) біоінформатикою.

**18. Наука, яка вивчає біологічні молекули, використовуючи методи прикладної математики, інформатики та статистики для аналізу інформації, пов'язаної з цими біологічними об'єктами називається .....**

- а) геномікою;
- б) транскриптомікою;
- в) протеомікою;
- г) метаболомікою;
- д) біоінформатикою.

**19. Доповніть відповідь: *Сукупність метаболітів живої системи називається .....***

**20. Виберіть положення, які відображають мету та завдання біоінформатики:**

- а) провести секвенування ДНК усіх живих організмів;
- б) впорядкувати всі наявні дані так, щоб вони були доступними для інших дослідників;
- в) розвивати програмні засоби та інформаційні ресурси, які дають змогу аналізувати наявні дані;
- г) вивчати протеом організмів за допомогою двовимірного електрофорезу;
- д) використовувати засоби та ресурси для аналізу інформації й інтерпретувати результати аналізу та висновки у необхідному для біології вигляді.



---

---

## ТЕМА 2. БІОІНФОРМАЦІЙНІ БАЗИ ДАНИХ

**Мета:** ознайомитися з різними видами біоінформаційних баз даних. Навчитися застосовувати бази даних для отримання та інтерпретації біологічної інформації.

### ТЕОРЕТИЧНА ЧАСТИНА

#### 1. Класифікація баз даних біологічної інформації

Дедалі більше біологічних даних публікуються не у звичайний спосіб, а вміщуються у бази даних із присвоєнням унікального ідентифікатора для посилання під час друку публікацій. Біологічні бази даних – це архіви узгоджених даних, які зберігаються в єдиній формі, тобто це сукупність пов'язаної інформації, об'єднаною за певними ознаками.

Біоінформаційні бази даних забезпечують зручне й ефективне зберігання великої кількості інформації, систематизацію амінокислотних і нуклеотидних послідовностей, спрямовану на їх порівняльний аналіз, зокрема для:

- транслювання амінокислотних послідовностей білків;
- ідентифікації організмів, їхньої таксономічної належності та рівнів еволюційного розвитку, побудови філогенетичних дерев;
- виявлення у неперервній послідовності символів окремих структурних одиниць та визначення їхніх функцій;
- розшифрування просторової структури білків;
- встановлення структурно-функціональних взаємозв'язків груп білків;
- виявлення генів, які кодують макромолекули-мішені для дії нових ліків та їхній синтез.

Основне призначення баз даних, крім збереження інформації – швидкий пошук і цілеспрямоване структурування відомостей. Вони також повинні забезпечувати користувача засобами для управління всіма їхніми даними та інструментами для аналізу відповідної інформації.

Дані від проектів зі секвенування геномів можуть навіть не мати посилань у журнальних публікаціях. Проте такі бази даних – дуже важливі інструменти для біологічних досліджень.

---

Бази даних із біологічною інформацією необхідні для більшості біоінформатичних досліджень. Є велика кількість баз, які містять усе – від нуклеотидних послідовностей до опису видів і фенотипів. Вони, як правило, доступні через інтернет і оснащені зрозумілим інтерфейсом для пошуку інформації. Багато з них перебувають у вільному доступі, інші закриті. Ці бази охоплюють дані широкого спектра різних галузей молекулярної біології.

Є кілька класифікацій баз даних. Із них розрізняють первинні та вторинні.

*Первинні*, або архівні бази даних містять анотовані первинні структури ДНК і білків, просторові структури нуклеїнових кислот і білків, а також профілі експресії генів білків клітин.

*Вторинні* зберігають результати аналізів первинних джерел, зокрема інформацію про специфічні мотиви в послідовностях, варіантах і мутаціях, а також еволюційних зв'язках. До цих же баз даних можна віднести і бібліографічні.

За інформацією, яка міститься в базах даних з біології, їх поділяють на:

- бібліографічні;
- первинних послідовностей ДНК, РНК, білків;
- просторової структури молекул;
- геномів;
- інші тематичні.

У світі є лише кілька великих центрів, які підтримують банки даних перших чотирьох типів, і велика кількість організацій, які містять бази даних з певних тем (наприклад, база даних із рибосомальних генів або класифікації ферментів; бази даних із мутацій гена p53 або білків *Saccharomyces cerevisiae*).

Є багато різних типів баз даних, які відрізняються за джерелом надходження інформації. Тому, залежно від організації інформації, бази даних поділяються на архівні, курирувані, похідні.

***Архівні бази даних:***

*GeneBank & EMBL* – тут зберігаються первинні послідовності всіх розшифрованих геномів;

*PDB* – база даних просторових структур білків.

---

---

### **Бази даних, які курируються:**

*UniProt* – найбільш якісна база даних, яка містить амінокислотні послідовності білків;

*KEGG* – містить інформацію про метаболічні шляхи;

*FlyBase* – охоплює інформацію про *Drosophila*;

*COG* – складається з даних про ортологічні гени, розміщених в алфавітному порядку.

### **Похідні бази даних**

Такі бази виникають унаслідок обробки даних з архівних і баз даних, які курируються:

*SCOP* – база даних структурної класифікації білків (описується структура білків);

*PFAM* – база даних про родини білків;

*GO* (Gene Ontology) – база даних із класифікації генів (здійснена спроба створення набору термінів, впорядкування термінології для того, щоб один і той же ген не називався по-різному і щоб різним генам не давали однакові назви);

*ProDom* – база даних білкових доменів.

Архівні бази даних належать до первинних і містять необроблені дані у тій формі, у якій вони були отримані із джерела. Бази даних, які курируються, належать до вторинних, які охоплюють відібрану після аналізу з архівних баз даних інформацію. Наприклад, GenBank – найбільша архівна база даних нуклеотидних послідовностей, а UniProt, яка курирується – найдостовірніша база даних про білки.

Ще в один тип баз даних можна виділити інтегровані, у яких зібрана вся інформація з певної тематики як із баз даних, які курируються, так і з баз даних, які не курируються. Якщо у таку базу даних ввести назву гена можна знайти все, що про нього відомо – у яких організмах він зустрічається, у якому місці локалізований, які функції виконує, яку просторову структуру має. Такою базою даних є NCBI Entrez (доступ до інформації про нуклеотидні й амінокислотні послідовності).

За тематикою бази даних часто поділяють на два типи, а саме на бази даних загального та спеціального призначень. Бази даних ДНК, білків, вуглеводів та інші є базами загального призначення. Як приклад спеціалізованих баз даних можна

---

---

навести базу даних поліморфізмів окремих нуклеотидів, послідовностей, які характеризують геном.

Щодо загального призначення розрізняють бази даних послідовностей і бази даних структур. Бази даних структур містять записи окремих послідовностей, структури макромолекул яких визначені біохімічними методами (наприклад база даних Protein 3D structure).

Інколи бази даних класифікують за тематикою (нуклеотидні послідовності або цілі геноми, амінокислотні послідовності або просторова структура білків, конкретні організми або наукова література). Розглянемо детальніше біоінформаційні бази даних, групуючи їх за тематиками.

## 2. Нуклеотидні бази даних

Прикладом вільних баз даних із інформацією про нуклеотидні послідовності є EMBL, GenBank та DDBJ, створені і підтримуються у рамках Міжнародної співпраці баз даних нуклеотидних послідовностей (International Nucleotide Sequence Database Collaboration).

**EMBL** (European Molecular Biology Laboratory) – <http://www.ebi.ac.uk/>.

EMBL – база даних усіх розшифрованих нуклеотидних послідовностей (ДНК і РНК) Європейської молекулярно-біологічної лабораторії, яка працює при Європейському інституті біоінформатики (the European Bioinformatics Institute (EBI)). База даних поповнюється переважно безпосередньо авторами, які визначили первинну структуру фрагмента ДНК або РНК, і, окрім послідовності нуклеотидів, містить інформацію про кожен фрагмент, зокрема літературні посилання та перехресні посилання на документи з інших баз даних.

База даних існує з 1982 року і є продуктом співпраці консорціуму, який складається з EMBL (Німеччина), GenBank (США) і DDJP (Японія), кожен з членів якого збирає свою порцію інформації зі всіх доступних джерел, щодня обмінюючись новими й оновленими документами. Зручна своєю географічною близькістю для доступу на території Європи.



## Explore EMBL-EBI and our mission

The European Bioinformatics Institute (EMBL-EBI) shares data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry. We are part of EMBL, Europe's flagship laboratory for the life sciences. [More about EMBL-EBI and our impact](#)

### Services

We provide freely available data and bioinformatics services to all facets of the scientific community

### Research

We contribute to the advancement of biology through basic investigator-driven research

### Training

We provide advanced bioinformatics training to scientists at all levels

### Industry

We help disseminate cutting-edge technologies to industry

### ELIXIR

We support, as an ELIXIR node, the coordination of biological data provision throughout Europe

**Рис. 2. Головний пошуковий екран EMBL**

EMBL належить до архівних баз даних. Це означає, що будь-яка заявка автора нового запису, після перевірки на відповідність форматам, вноситься до бази даних. Куратор бази контролює тільки, щоб не було очевидних неточностей. Кожен запис у банку даних має унікальний ідентифікатор.

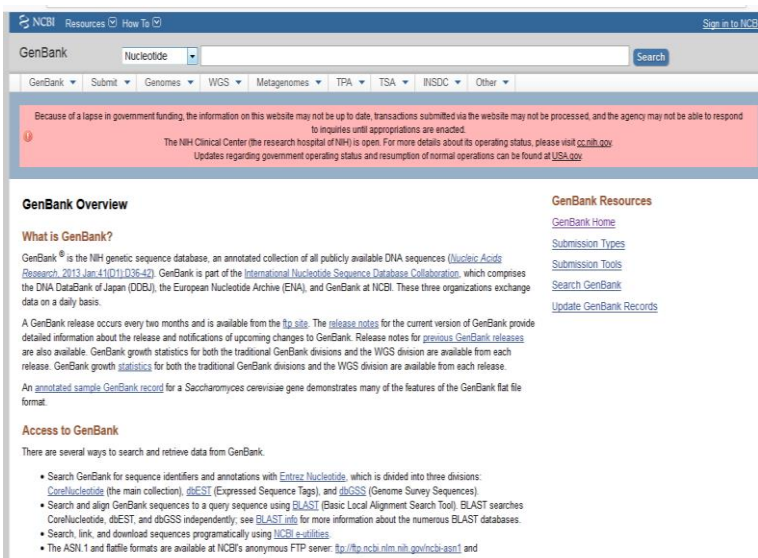
Основна структурна одиниця – нуклеотидна послідовність ділянки ДНК чи РНК (можливо вся РНК) з клітини організму чи іншого джерела. Окрім геномів та нуклеотидних послідовностей, в базі даних можна знайти інформацію про: білкові послідовності; макромолекулярні структури та біоактивні молекули; молекулярні взаємодії; хвороби тощо.

Нині EMBL складається з 18 розділів, більша частина яких відображає таксономію. Поточний реліз EMBL містить 508708760 записів ДНК-послідовностей та 53467764 записів РНК-послідовностей відповідно до індексації від 24 січня 2018 року.

Отже, EMBL підтримується EBI у співпраці з the DNA Data Bank of Japan (DDBJ) та GenBank у NCBI (National Center for Biotechnology Information, USA).

**GenBank** – <http://www.ncbi.nlm.nih.gov/Genbank>.

GenBank – найвідоміша база даних при Національному центрі біотехнологічної інформації (США).



**Рис. 3. Головний пошуковий екран GenBank**

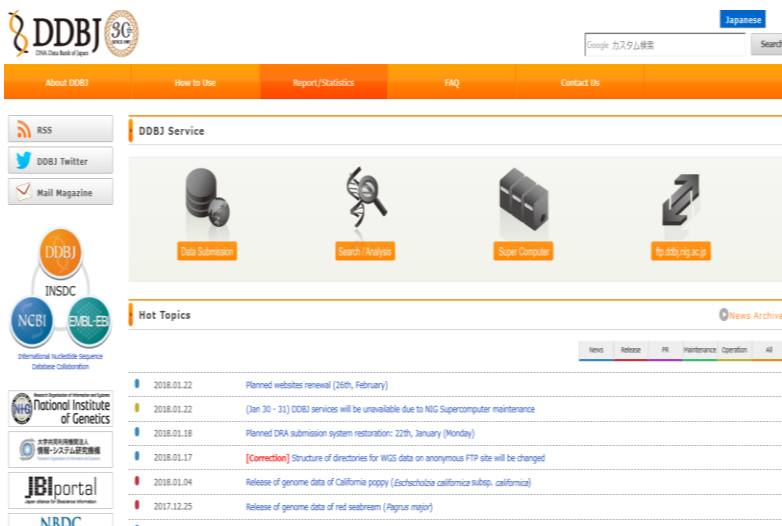
Ця база даних нуклеотидних (ДНК/РНК) та білкових послідовностей заснована в 1982 році та пропонує безкоштовний публічний доступ до майже всіх послідовностей ДНК/РНК та білків, які транскрибуються з них.

GenBank містить анотовану колекцію всіх загально-доступних послідовностей ДНК, РНК та білків разом із літературними посиланнями. Послідовності нуклеїнових кислот подані як у графічному зображенні, так і у форматі FASTA. Поповнюється раз на два місяці. Станом на жовтень 1999 р. містила 8293265 записів нуклеотидних послідовностей, на лютий 2004 р. – 32549400 записів послідовностей. Відповідно до

індексації від 24 січня 2018 року в базі даних нараховується 124385078 нуклеотидних послідовностей.

Базу формують у співпраці з Європейською лабораторією молекулярної біології (EMBL) та Японською базою даних ДНК (DDBJ), тому вона щоденно синхронізується з DDBJ та EMBL.

**DDBJ** (Center for Information Biology (CIB), DNA Data Bank of Japan) – <http://www.ddbj.nig.ac.jp>.



**Рис. 4. Головний пошуковий екран DDBJ**

DDBJ – база даних генетичних послідовностей. Збір інформації виконується насамперед із японських учених та літератури. 75 % зібраних послідовностей являють собою частково секвеновані фрагменти ДНК із кількох сотень експресованих генів, так званих EST (Expressed Sequence Tags). База даних заснована в 1984 році.

EMBL, GenBank, DDBJ – усі три бази містять однакову інформацію, розбіжності пов'язані тільки з часом внесення запису, який заново надходить. Ці бази даних обмінюються інформацією щодня, внаслідок чого складаються з майже ідентичної інформації. Ці нуклеотидні бази даних приймають інформацію про послідовності ДНК і надають відкритий доступ

---

---

до неї. Доступ до цих баз даних безкоштовний і можливий через Інтернет.

### **Інші нуклеотидні бази даних**

**UniGene** (<http://www.ncbi.nlm.nih.gov/UniGene>) – база даних, яка містить кластери схожих послідовностей. Кожен кластер – це один ген. У базі даних також міститься супутня інформація, наприклад, назва тканини, в якій цей ген експресований. UniGene поєднує близькі сіквенси. Окрім добре відомих генів, у базу даних внесені сотні тисяч нових кінців послідовностей, які експресуються. Служить для пошуку генів в нових послідовностях. Кластеризація здійснюється автоматично.

**STACK** (Sequence Tag Alignment and Consensus Knowledgebase) (<http://www.sanbi.ac.za/Dbases.html>) – база даних подібна до попередньої.

**EMBL-SVA** (EMBL-Sequence Version Archive) (<http://www.ebi.ac.uk/embl/sva>) – містить усі записи з першого випуску бази даних EMBL (понад 100 млн записів).

**Геномні бази даних** значно різняться за формою і вмістом.

Для найважливіших і цікавих із позицій генетики організмів є опубліковані каталоги генів і мутацій у них. Останніми роками багато каталогів переведені в електронну форму.

Крім того, створено чимало нових баз генів, які відрізняються за доборою і способом подання інформації.

**Genomes Server** (<http://www.ebi.ac.uk/genomes>) – база даних надає доступ до великої кількості повних геномів різних організмів.

**Proteome Analysis** (<http://www.ebi.ac.uk/proteome/index.html>) – база даних, яка спирається на статистичний і порівняльний аналіз передбачених протеомів повністю секвенованих організмів.

**HGMD** (Human Gene Mutation Database) (<http://archive.uwcm.ac.uk/uwcm/mg/hgmd.html>) – містить інформацію про всі опубліковані пошкодження генів, які призводять до спадкових захворювань у людини. Документи бази анотують всі гени, розміщені в ядрі. Гени мітохондріального генома і соматичні мутації вилучені. Мутації, виявлені на рівні



---

білкового секвенування, не внесені в базу, щоб уникнути помилок аналізу на рівні ДНК. Мутації, які «мовчать» і не призводять до зміни амінокислотних послідовностей, теж вилучені. Із березня 1999 року внесені дані про поліморфізм, пов'язаний із хворобами. Дані беруться з тих самих журналів, що і дані про мутації. Супроводжується Інститутом медичної генетики (University of Wales Cardiff, UK).

**GAD** (Genetic Associated Database) (<http://geneticassociationdb.nih.gov>)

– база даних, яка містить архів досліджень зв'язків між генетичними порушеннями та хворобами. Її призначення – можливість швидкої ідентифікації поліморфізмів, які мають значення в медицині у великому обсязі інформації із поліморфізмів і мутацій.

**NDB** (Nucleic Acids Database) (<http://ndbserver.rutgers.edu>) –

база даних, яка містить інформацію про нуклеїнові кислоти різних організмів.

**Ensembl** (<http://www.ebi.ac.uk/ensembl/index.html>) – спільний

проект ЕБІ и Wellcome Trust Sanger Institute, створений для розробки системи, яка підтримує автоматичне анотування великих еукаріотичних геномів. Являє собою постійно оновлювану базу з автоматичним анотуванням геномів багатоклітинних організмів. Доступні людина, миша, щур, риба фугу, комар, дрозофіла, нематоди.

**Karyn's Genomes** (<http://www.ebi.ac.uk/2can/genomes/index.html>)

– надає загальну інформацію про організми, чий геном повністю секвенований.

**WormBase** (<http://www.wormbase.org>) – банк інформації щодо

картування, секвенування і фенотипів *C. elegans* та деяких інших нематод.

**FlyBase** (<http://flybase.bio.indiana.edu>) – база для *Drosophila*

*melanogaster*.

**MGD** (the Mouse Genome Database) (<http://www.informatics.jax.org>) –

база даних із геномів миші.

**RGD** (Rat Genome Database) (<http://rgd.mcw.edu>) – база даних із

геномів щура.

**SGD** (*Saccharomyces* Genome Database) (<http://www.stanford.edu/Saccharomyces>) – база даних, яка вміщує інформацію про

геноми дріжджів.

- 
- MIPS** (<http://www.mips.biochem.mpg.de/proj/yeast>) – база із генома дріжджів.
- SPGP** (S. Pombe Genome Project) ([http://www.sanger.ac.uk/Projects/S\\_pombe](http://www.sanger.ac.uk/Projects/S_pombe)) – база про грибки *Schizosaccharomyces pombe*.
- AceDB** (<http://www.acedb.org>) – база, яка містить генетичну інформацію *Caenorhabditis elegans*. Система управління нею дуже популярна і лягла в основу багатьох інших баз даних. Однойменна назва бази і системи управління базою ACEDB призвели до деякої плутанини серед баз даних із *C.elegans*.
- HIV-SD** (HIV Sequence Database) (<http://hiv-web.lanl.gov/content/hivdb/mainpage.html>) – збирає, курує та анотує ДНК-послідовності вірусу імунодефіциту людини (human immunodeficiency virus, HIV) та вірусу імунодефіциту мавпи (simian immunodeficiency virus, SIV).

#### Бази даних геномів рослин

- MaizeGDB** (Maize Genetics and Genomics Database) (<https://www.maizegdb.org/person?id=16635>) – база даних із геномів рослин.
- PGDIC** (Plant Genome Information Resource) (<http://www.nal.usda.gov/pgdic>) – доступ до баз геномів багатьох рослин, зокрема бавовни, люцерни, пшениці, ячменю, жита, рису, проса, сорго, пасльонових і дерев.
- MENDEL** (<http://www.mendel.ac.uk>) – широка база даних із геномів рослин.

Отже, найбільші бази даних нуклеотидних послідовностей – це EMBL, GenBank, DDBJ, які поповнюються дослідниками через Інтернет. Надалі ці дані проходять перевірку. Іншим джерелом інформації для них є наукова література. Більшість баз даних із подібною біологічною інформацією постійно синхронізуються, наприклад бази даних нуклеотидних послідовностей EMBL, GenBank, DDBJ.

Окрім того, існує можливість описати фізіологію організму за допомогою комп'ютерного аналізу його генома, шляхом знаходження подібних структурних одиниць у геномах вже вивчених організмів. В даний час найефективнішим методом

визначення біологічної функції гена є пошук схожих на нього нуклеотидних послідовностей у наявних базах даних.

### 3. Бази даних білків і амінокислотних послідовностей

Бази даних білків і амінокислотних послідовностей можна класифікувати як первинні, змішані та вторинні. Первинні бази даних містять понад 300 тисяч послідовностей білків і функціонують як «сховища» для необробленої інформації. Проте деякі первинні бази даних, такі як SWISS-PROT і PIR-International, анують послідовності, пов'язуючи їх із функціями білків, доменною структурою та посттрансляційними модифікаціями.

Є універсальні бази даних і спеціалізовані, які покривають вузькі родини чи групи білків, або білки з певного організму.

Один із найбільших каталогів інформації про білок і його функції, а також центральне сховище первинних структур – база даних UniProt. У ній містяться дані про білки більш як 6000 тисяч організмів.

**UniProtKB** (Universal Protein Resource Knowledgebase) (<http://www.uniprot.org>) – база даних послідовностей білків. Інформація з бази даних доступна для всіх користувачів і охоплює різні аспекти аналізу білкових послідовностей.

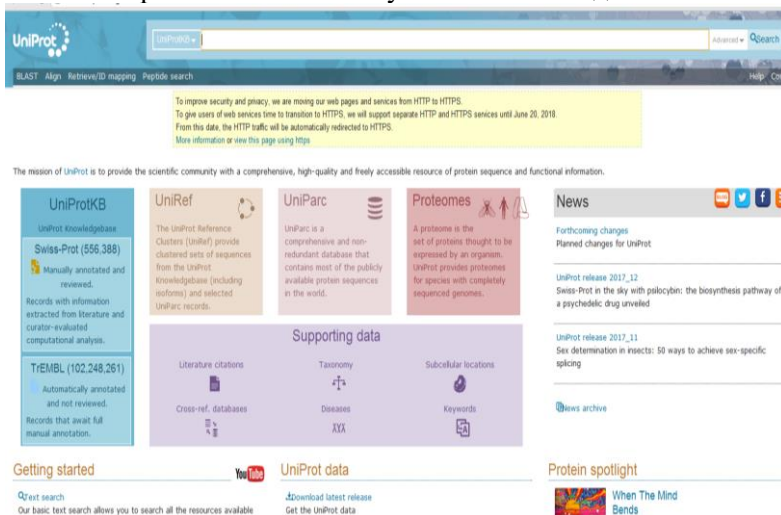


Рис. 5. Головний пошуковий екран UniProt

---

Багато білкових послідовностей стали відомими внаслідок реалізації проектів секвенування геномів останніх років. Крім того, база даних UniProt містить велику кількість інформації про біологічні функції білків, отриманої з наукової літератури.

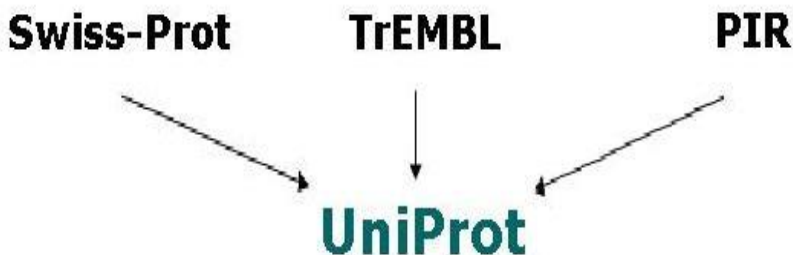
До UniProt-консорціуму належать: Європейський інститут біоінформатики (EBI), Швейцарський інститут біоінформатики (SIB) і Білковий інформаційний ресурс (PIR).

У EBI, розташованому у Великій Британії, розміщена велика кількість біоінформативних баз даних і серверів. SIB (Женева, Швейцарія), є сховищем серверів для експертного білкового системного аналізу (ExPASy-серверів), що є основним джерелом для інструментів протеоміки та відповідних баз даних. PIR (Медичний центр університету Джорджтауна у Вашингтоні, США) – інтегрований біоінформативний ресурс, призначений для підтримки досліджень у геноміці і протеоміці.

У 2002 році PIR, спільно зі своїми міжнародними партнерами, EBI і SIB, отримали грант від Національного інституту здоров'я для створення UniProt, єдиної всесвітньої бази даних послідовностей і функцій білків. Так з'явився консорціум UniProt. Проект UniProt почав діяти з грудня 2003 року.

UniProt фінансується за рахунок грантів від Національного інституту дослідження генома людини (NHGRI), Національного інституту здоров'я (NIH), Європейської комісії, Швейцарського федерального уряду через Федеральне управління освіти і науки, Національного інституту раку і ракової біомедичної інформаційної мережі (NCI-caBIG).

Єдина база даних UniProt створена об'єднанням баз даних Swiss-Prot, TrEMBL і PIR.



---

Swiss-Prot (Swiss Protein Databank) створена в 1986 році Амосом Байрошем під час роботи над своїм PhD-проектом і розвинена в подальшому в Швейцарському інституті біоінформатики (SIB), а згодом допрацьована Рольфом Апвейлером в EBI. Основна функція бази даних Swiss-Prot спрямована на забезпечення надійності інформації про білкові послідовності, обумовленої високим, детальним рівнем анотації, виконаної вручну. Вона охоплює опис функції білка, його доменної структури, посттрансляційних модифікацій, різних варіантів послідовності, причому з високим рівнем інтеграції з іншими базами даних.

TrEMBL (Translation of EMBL Nucleotide Sequence Database) розроблена в 1996 році як анотований комп'ютерний додаток до Swiss-Prot. Рішення про створення TrEMBL схвалене у відповідь на збільшення потоку даних у результаті появи проектів геномів, а витратний за часом і трудомісткий процес ручної анотації в Swiss-Prot перевищував можливості Swiss-Prot для того, щоб внести всі доступні білкові послідовності. TrEMBL надає можливість автоматизованої анотації для трансляції наявних нуклеотидних послідовностей і перетворення їх на білкові послідовності зовні Swiss-Prot.

TrEMBL складається з амінокислотних послідовностей, отриманих трансляцією всіх кодувальних послідовностей ДНК з бази EMBL за винятком тих, які уже внесені до Swiss-Prot.

TrEMBL складається з двох основних секцій:

SP-TrEMBL містить записи, які будуть внесені до Swiss-Prot.

REM-TrEMBL (REMAining TrEMBL) містить послідовності, отримані синтетично, неповні, є псевдогенами, імуноглобулінами або Т-клітинними рецепторами. Вони не цікаві для анотування і не внесені до бази даних Swiss-Prot.

PIR (Protein Information Resource) організований Національним фондом медико-біологічних досліджень (NBRF) у Медичному центрі університету Джорджтауна у Вашингтоні в 1984; публікується як «Atlas of Protein Sequence and Structure».

PIR – спадкоємець старої бази даних послідовностей білків, а саме, створеним Margaret Dayhoff «Атласом послідовностей білка і структури», вперше опублікованим у 1965 році. З 1988 року – це міжнародна база даних.

---

PIR підтримує кілька білкових баз даних, а саме: головну базу білкових послідовностей (PIR-PSD), базу даних, пов'язану з класифікацією білків за структурою та функціями (iProClass), а також інші бази даних білкових послідовностей.

База даних поділена на чотири секції: PIR1, PIR2, PIR3 і PIR4.

PIR1 повністю класифікована за суперродинами й анотована.

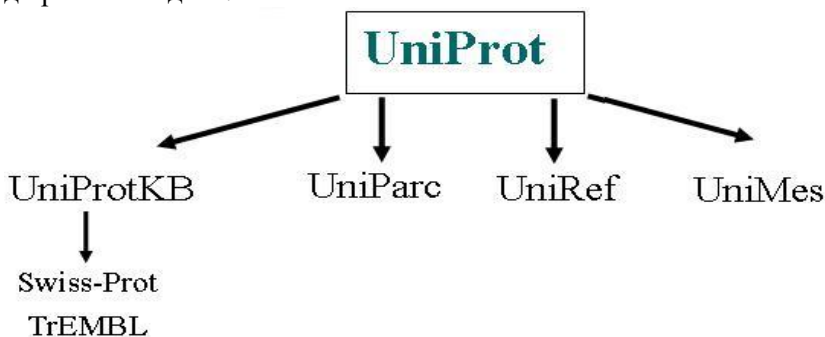
PIR2 – перехідний розділ від PIR3 до PIR1.

PIR3 слугує тимчасовим розділом для нових надходжень.

PIR4 містить некласифіковані послідовності.

Після об'єднання кожен член консорціуму активно бере участь у забезпеченні єдиної бази даних послідовностями білків і їхній анотації.

UniProt складається з чотирьох компонентів оптимізованих до різних завдань:



UniProt (The UniProt Knowledgebase) – центр доступу, зокрема до функцій, класифікації та перехресних посилань.

UniParc (The UniProt Archive) – всеосяжна база, яка відображає історію всіх первинних структур білків.

UniRef (The UniProt Reference Clusters) – комбінує близькі структури в одному записі для прискорення обробки.

UniMes (The UniProt Metagenomic and Environmental Sequences database) – в даний час містить дані про білкові послідовності організмів зі світового океану, забезпечених глобальною океанічною експедицією зі збору проб.

---

Основна структурна одиниця банку UniProt – амінокислотна послідовність нативного білка чи фрагмента такої послідовності; химери (створені руками людини) не внесені.

UniProt – база даних, яка перебуває під наглядом кураторів-експертів. Це означає, що заявка автора нової послідовності потрапляє до рук куратора бази даних, який є експертом із відповідної родини білків. Куратор не тільки перевіряє те, що написав автор про запропонований білок, але й коректує цю інформацію, додає нову, отриману зі скринінгу кількох баз даних, оцінює достовірність опису, використовуючи прийняту в UniProt термінологію. Така технологія поповнення допомагає значно поліпшити описи послідовностей і зробити їх більш формалізованими, зменшити число помилок у базі даних.

### **Бази даних структури білків**

**PDB** (Protein Data Bank) (<https://www.rcsb.org>) – банк даних 3D-структур білків. Інформація, отримана методами рентгенівської кристалографії або ЯМР-спектроскопії, вноситься до бази даних біологами та біохіміками зі всього світу, і доступна безкоштовно через Інтернет.

PDB – один із найважливіших ресурсів для учених, які працюють у галузі структурної біології. Більшість наукових журналів і деякі фонди фінансування досліджень у США вимагають від авторів статей і одержувачів грантів, щоб усі структурні дані були внесені до PDB. PDB містить, в основному, первинні дані про структуру біологічних молекул, тоді як існують сотні інших банків даних, які категоризують первинні дані або закономірності, виявлені між будовою молекул і еволюційною спорідненістю.

PDB створений у 1971 р. Волтером Хемілтоном у Національній лабораторії Брукхавена. У жовтні 1998 року PDB перенесений в Research Collaboratory for Structural Bioinformatics (RCSB); перенесення інформації завершено в червні 1999 року.

Окрім RCSB, засновниками wwPDB є Protein Data Bank in Europe (PDBe), Protein Data Bank in Japan (PDBj) і Biological Magnetic Resonance Bank (BMRB).



**Рис. 6. Головний пошуковий екран PDB**

База даних PDB оновлюється щотижня (о 0:00 годині за всесвітнім координованим часом). Станом на 2018 рік вона містить близько 42065 різних білкових послідовностей, 37366 структур із послідовностями людини і 9789 комплексів білків із нуклеїновими кислотами. Кожна структура, опублікована в PDB, отримує чотиризначний ідентифікатор (комбінація цифр і літер латинського алфавіту) (наприклад 1A4E – каталаза з *Saccharomyces cerevisiae*). Цей шифр не може служити ідентифікатором біомолекул, оскільки часто різні структури однієї і тієї ж молекули, наприклад у різному середовищі, можуть мати різні PDB-ID.

Є величезна кількість спеціалізованих білкових баз даних. Вони цінні тим, що як правило доповнюють бази первинних структур додатковою аналітичною інформацією.

#### **Спеціалізовані бази даних структури білків**

**SCOP** (Structural Classification of Proteins) (<http://scop.mrc-lmb.cam.ac.uk/scop>) – база даних структурної класифікації білків, створена за допомогою ручної обробки наявних просторових структур білкових молекул (PDB).





Welcome to SCOP: Structural Classification of Proteins.  
1.75 release (June 2009)

38221 PDB Entries, 1 Literature Reference, 110800 Domains (excluding nucleic acids and theoretical models)

Folds, superfamilies, and families [analytics here](#)

[New folds superfamilies families](#)

[List of obsolete entries and their replacements](#)

**Authors:** Alexey G. Morin, John-Marc Chandonia, Antonia Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Alely, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)

**Reference:** Morin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]

**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Morin A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* 30(1), 264-267. [PDF]

Andreeva A., Howorth D., Brenner S. E., Hubbard T.J.P., Chothia C., Morin A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.* 32(D126-D129) [PDF]

Andreeva A., Howorth D., Chandonia J.-M., Brenner S. E., Hubbard T.J.P., Chothia C., Morin A. G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* advance access. doi:10.1093/nar/gkm993 [PDF]



The prototype of a new Structural Classification of Proteins 2 (SCOP2) database is now available at <http://scop2.mrc-lmb.cam.ac.uk/>

#### Access methods

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)

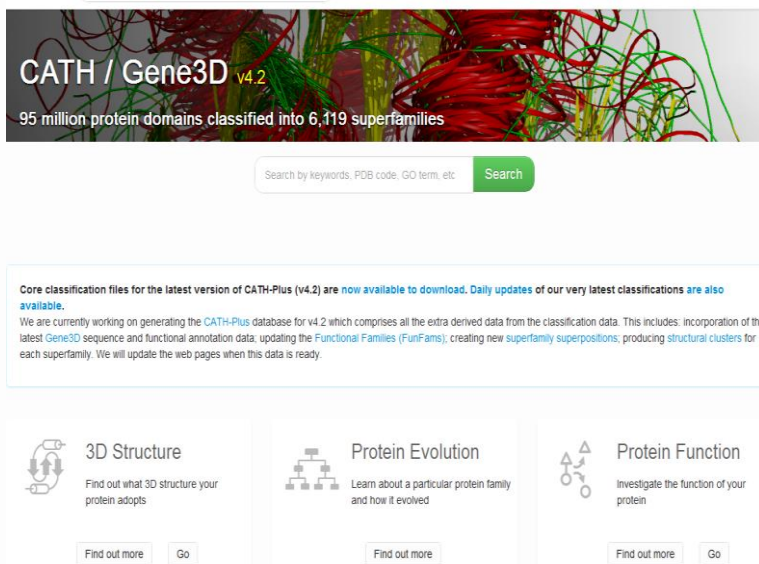
### Рис. 7. Головний пошуковий екран SCOP

Координатори бази даних комплексно впорядковують усі білки з відомою структурою, основується на еволюційних і структурних взаємодіях. Має ієрархічну організацію. База даних спирається на експертні знання. SCOP поділяє домени на класи (4 основні –  $\alpha$ ,  $\beta$ ,  $\alpha/\beta$ ,  $\alpha+\beta$ ), фолди, суперродини та родини. Фолди (folds) поєднують білки з однаковою організацією вторинної структури, схожим просторовим розташуванням і зв'язками; білки, об'єднані одним фолдом, можуть не мати загального предка. У суперродини (superfamily) об'єднують білки, які мають низьку схожість послідовностей, але їхня структурна і функціональна подібність дає змогу передбачити наявність загального предка. Родини (family) об'єднують білки з очевидними еволюційними взаємозв'язками (парна подібність послідовностей >30 %).

**CATH** (Class, Architecture, Topology and Homology) (<http://www.cathdb.info>) – база даних структурної класифікації білків, основана на автоматичній обробці інформації (але інспектується вручну).

Основні елементи ієрархії CATH, як випливає з назви, – це клас, архітектура, топологія, гомологія. Три основні класи  $\alpha$ ,  $\beta$  і  $\alpha/\beta$ . Архітектура об'єднує домени за загальним виглядом структури, без урахування зв'язків. Топологія об'єднує структури з урахуванням загального вигляду та зв'язків елементів вторинної структури. Наступний рівень об'єднує

структури на основі можливого загального предка (гомології), далі вони також можуть об'єднуватися на основі ідентичності послідовностей. Акцент робиться на структурну класифікацію.



**Рис. 8. Головний пошуковий екран CATH**

SCOP і CATH по-різному виділяють класи, мають різні способи класифікації. SCOP – акцентує еволюційні взаємозв'язки, а CATH – структурну класифікацію. Це призводить до різних результатів класифікації одних і тих же білків.

**PROSITE** (Database of protein domains, families and functional sites) (<https://prosite.expasy.org>) – база даних, яка містить інформацію про родини білків, виділених на основі ідентичності специфічних сайтів усередині кожної родини. Опис кожної специфічної ділянки (сайта) кожного з подібних родин (разом із прикладами, взятими з бази даних SWISS-PROT) становить один запис.

PROSITE складається з документів, які описують білкові домени, родини та функціональні сайти, а також пов'язані з ними шаблони та профілі для їхньої ідентифікації.



## Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users]. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2017\_12 of 20-Dec-2017 contains 1797 documentation entries, 1309 patterns, 1202 profiles and 1223 ProRule.

**Рис. 9. Головний пошуковий екран PROSITE**

PROSITE доповнюється ProRule, набір правил на основі профілів і патернів, які підвищують дискримінаційну здатність профілів і моделей, надаючи додаткову інформацію про функції та структуру амінокислот. Станом на кінець грудня 2017 року база даних містила 1797 записів, 1309 патернів, 1202 профілів і 1223 ProRule.

### **Спеціалізовані бази даних білків**

На основі інформації про амінокислотні послідовності є спеціалізовані бази даних, які містять інформацію про окремі класи сполук.

**ENZYME** (<http://www.expasy.ch/enzyme>) – анотоване розширення UniProtKB/Swiss-Prot із ферментів. Містить інформацію щодо номенклатури ферментів і описує всі типи білків, яким присвоєно номер EC (Enzyme Commission). Пошук реалізовано за EC-номером, класами ферментів, хімічними компонентами, кофакторами, назвами хвороб, пов'язаних із ферментами.

**ExpASy** (<https://www.expasy.org>) – ресурсний портал біоінформатики SIB, який надає доступ до наукових баз даних та програмних засобів (тобто ресурсів) у різних

---

---

галузях наук про життя, зокрема протеоміки, геноміки, філогенезії, системної біології, популяційної генетики, транскриптоміки (меню відображається на головному пошуковому екрані ліворуч). База даних спеціалізується на протеоміці та є найбільшим банком із двовимірних гелів (<http://www.expasy.ch/ch2d>).

**MEROPS** (<https://www.ebi.ac.uk/merops>) – каталог і структурна класифікація пептидаз у родини на основі «пептидазних одиниць».

**BRENDA** (<http://brenda-enzymes.info>) – база даних із властивостей ферментів.

**GPCRDb** (<http://gpcrdb.org>) – база із G-білкових рецепторів, великої родини білків, які є важливими компонентами різних сигнальних систем тварин.

**YPD** (the yeast protein database) (<https://www.yeastgenome.org>) – база для білків із *S. cerevisiae*.

Існують бази даних протеїнкіназ, протеїнази ВІЛ, імуноглобулінів, рецепторів Т-лімфоцитів.

#### 4. Бази даних метаболічних шляхів

Найдостовірнішу інформацію про метаболічні шляхи різних організмів можна отримати на ресурсі KEGG.

**KEGG** ([www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)) – це веб-сервер-ресурс, який поєднує низку біологічних баз даних, у якому зібраний геном, хімічна, функціональна й інша інформація, і призначений, найперше для інтерпретації даних секвенування генома. Ресурс – це спроба комп'ютеризувати всі дані молекулярної та клітинної біології.

Сервер KEGG створений у 1995 році професором Кіотського університету в рамках японського проекту «Геном людини». Найпершою виникла база даних KEGG PATHWAY, яка є набором карт метаболічних шляхів, котрі відображають експериментальні дані про метаболізм і ферменти послані на гени. Завдяки цьому можна виконати аналіз, названий «pathway mapping». Через обмеження державного фінансування проекту з липня 2011 року для скачування баз даних по FTP потрібна передплата.



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

Menu PATHWAY BRITE MODULE KO GENOME GENES LIGAND DISEASE DRUG DBGET

Select prefix:

map

Organism

Enter keywords:

Go

Help

[ [New pathway maps](#) | [Update history](#) ]

### Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

#### 1. Metabolism

Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan  
Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical  
structure

#### 2. Genetic Information Processing

#### 3. Environmental Information Processing

#### 4. Cellular Processes

#### 5. Organismal Systems

#### 6. Human Diseases

and also on the structure relationships (KEGG drug structure maps) in:

#### 7. Drug Development

### Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretation of higher-level systemic functions.

- Search Pathway - basic pathway mapping tool
- Search&Color Pathway - advanced pathway mapping tool
- Color Pathway - selected pathway map coloring tool

## Рис. 10. Головний пошуковий екран KEGG PATHWAY

Ресурс підтримується Інститутом хімічних досліджень Кіотського університету в Японії. Набір інструментів для пошуку в базах даних KEGG має назву KEGG Mapper і містить низку баз-підрозділів: бази даних метаболічних шляхів (PATHWAY), генів (GENES), лігандів (LIGAND) (див. табл. 1).

Таблиця 1

### Бази даних веб-ресурсу KEGG

Інформація	Бази даних	Вміст
системна інформація	KEGG PATHWAY	схеми метаболічних і регуляторних шляхів
	KEGG BRITE	класифікація біологічних об'єктів і явищ
	KEGG MODULE	функціональні одиниці генів
геноми	KEGG ORTHOLOGY	родини ортологічних генів
	KEGG GENOME	секвовані геноми
	KEGG GENES	гени та білки з відсеквованих геномів

<i>хімічні сполуки</i>  (KEGG LIGAND)	KEGG COMPOUND	метаболіти, хімічні сполуки.
	KEGG GLYCAN	полісахариди
	KEGG REACTION	хімічні реакції
	KEGG RPAIR	пари хімічних реакцій
	KEGG RCLASS	класи хімічних реакцій
	KEGG ENZYME	номенклатура ферментів
<i>здоров'я людини</i>  (KEGG MEDICUS)	KEGG DISEASE	захворювання людини
	KEGG DRUG	лікарські препарати
	KEGG ENVIRON	різні хімічні сполуки
	JAPIC	лікарські препарати в Японії
	DailyMed	лікарські препарати в США

Карти метаболічних шляхів KEGG, функціональні ієрархії BRITE і модулі KEGG створюються та курируються експертами вручну.

За словами розробників інтеграція генів, білків, малих молекул, реакцій, взаємодій, KEGG – це «електронний представник» біологічних систем.

Важлива особливість KEGG – це факт, що всі бази даних всередині ресурсу пов'язані мережею перехресних посилань, чим і пояснюється інтерес до цього ресурсу.

Бази даних в KEGG подані у вигляді 3 граф (уявлення даних про гени, хімічні сполуки, білки і їх взаємодії). Графи складаються з вузлів (будівельних блоків) і ребер (взаємодій). Графи для генів містять інформацію про гомологічність генів і інші взаємозв'язки в секвенованих геномах. Графи для хімічних сполук показують хімічні та структурні взаємодії між біохімічними сполуками. У графі для білків наводиться інформація про відомі молекулярні зв'язки білків в різних клітинних процесах.

**EMP** (Enzymes and Metabolic Pathway) (<http://emp.mcs.anl.gov>) – база даних про ферменти та метаболічні шляхи. Містить

---

---

понад 10 000 оригінальних публікацій із ензимології та обміну речовин. У базі даних відображено понад 1800 метаболічних шляхів, вільно доступних у World Wide Web.

#### **Інші спеціалізовані бази даних**

**RDP** (the Ribosomal Database Project) (<http://rdp.cme.msu.edu>) – забезпечує інформацією щодо рибосом, включаючи аналіз даних онлайн, філогенетичні дерева із рРНК, які бідуються вирівнюванням анотованих послідовностей рРНК.

**IMGT** (ImMunoGeneTics database) (<http://www.imgt.org>) – спеціалізується на імуноглобулінах, Т-клітинних рецепторах і головному комплексі гістосумісності всіх видів хребетних.

**TRANSFAC** (<http://genexplain.com/transfac/>) містить дані із ДНК транскрипційних факторів еукаріот і ділянок, які їх пов'язують.

**EPD** (Eukaryotic Promoter Database) (<http://epd.vitalit.ch/index.php>) – анотована колекція еукаріотичних промоторів полімераза, для яких експериментально визначений стартовий сайт транскрипції.

**SRS** (Sequence Retrieval System) (<http://srs.ebi.ac.uk/>) – досить потужна система запитів, яка існує при Європейському біоінформаційному інституті (EBI). Забезпечує інформацією щодо понад 150 гетерогенних джерел.

**REBASE** (Restriction Enzyme Database) (<http://rebase.neb.com/rebase/rebase.html>) – база із ферментів і сайтів рестрикції.

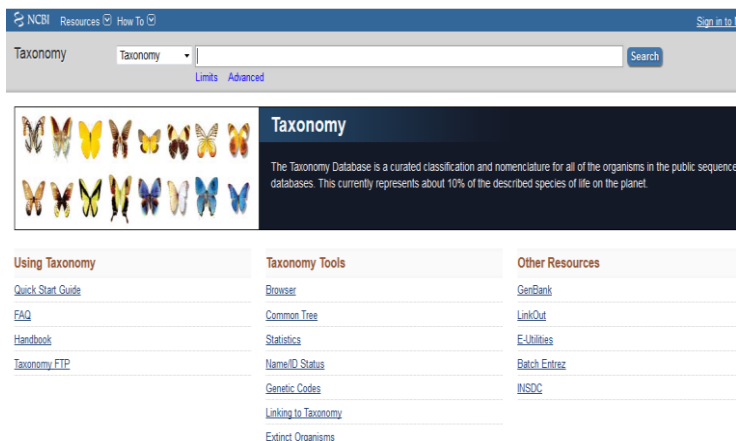
**GOBASE** (Organelle Genome Database) (<http://gobase.bcm.umontreal.ca>) – спеціалізована база геномів органел.

### **5. Таксономічні бази даних**

Таксономічні бази даних створені для класифікації всіх організмів. Мета цих баз даних – централізувати класифікацію всіх організмів, поданих у відповідній базі даних. Їх використовують для визначення положення досліджуваного організму в ієрархії, для здобуття послідовностей генів цього організму або групи організмів.

**Taxonomy** (<http://www.ncbi.nlm.nih.gov/Taxonomy>) – найпопулярніша ієрархічна база даних, розміщена при NCBI і основана на нуклеотидних послідовностях генів.

Проект Таксоному NCBI започаткований у 1991 році, коли проектувалася перша версія пошукової системи Entrez. База даних містить імена організмів і таксономічне положення для кожної із послідовностей поданих у базах даних нуклеотидних і білкових послідовностей в INSDC (International Nucleotide Sequence Database Collaboration).



*Рис. 11. Головний пошуковий екран Таксоному*

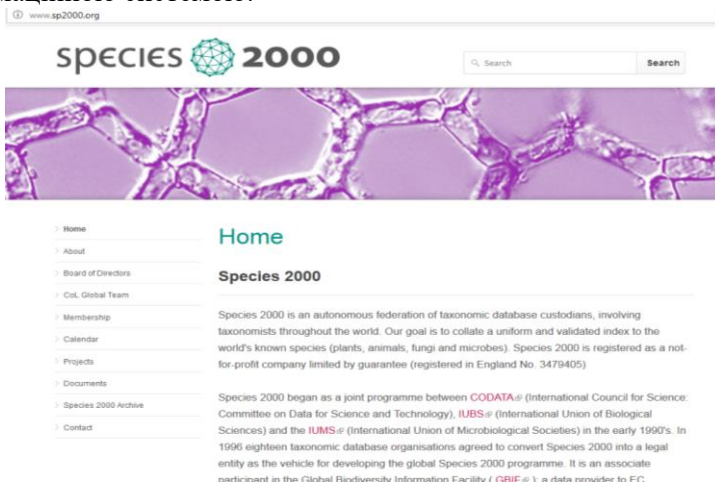
База даних Таксоному курирується вручну невеликою групою вчених в NCBI, які використовують таксономічну літературу та філогенетичний аналіз для вихідних організмів, представлених у базах даних послідовностей. База даних Таксоному – центральний організаційний центр для багатьох ресурсів NCBI.

***The Tree of Life project*** (<http://tolweb.org/tree/phylogeny.html>)

Веб-проект Tree of Life («Дерево життя») – це набір інформації про біологічне різноманіття, складений спільно сотнями експертів та аматорів. Ресурс містить сторінки із зображеннями, текстом та іншою інформацією для кожного виду та групи організмів, які живуть або вимерли. Зв'язки між веб-сторінками «Дерево життя» підпорядковуються філогенетичним розгалуженням між групами організмів, тому відвідувачі можуть переглядати ієрархію життя та дізнатися про філогенез та еволюцію, а також характеристики окремих груп.



**Species 2000** (<http://www.sp2000.org>) – таксономічна база даних. Species 2000 – член Глобального інформаційного фонду з питань біорізноманіття та робочої групи таксономічних баз даних. Співпрацює з інтегрованою таксономічною інформаційною системою.



**Рис. 12. Головний пошуковий екран Species 2000**

Species 2000 регулюється радою директорів, яка відповідає за правові та фінансові питання.

**NEWT** (<http://www.ebi.ac.uk/newt>) – новий таксономічний портал SWISS-PROT, який містить таксономічні дані, оновлювані щодня.

**DPD** (The Database of Plant Databases) (<http://www.biologie.uni-hamburg.de/b-online/ibc99/iopi/iopihome.html>) – база даних, яка містить інформацію про вищі рослини: таксономічну інформацію щодо видів рослин; каталоги гербаріїв.

**ITIS** (Integrated Taxonomic Information System) (<http://www.itis.gov>).

База даних створена всередині 1990-х як кооперативний проект кількох федеральних агентств із метою поліпшення і розширення таксономічних даних, які обслуговуються національним океанографічним центром.

ITIS містить близько 210000 наукових імен. Однак багато таксономічних груп не дуже добре подані (наприклад земні

---

комахи), деякі помилки й упущення трапляються в межах таксономічних груп (наприклад кілька неправильних написань або випадкових друкарських помилок, багато назв видів без авторів). Мета ITIS – створити наукову базу даних таксономічної інформації.

Таксономічні бази даних достатньо суперечливі через різницю у поглядах на класифікацію організмів.

## 6. Бази даних наукової інформації

Бази даних літератури містять бібліографічні дані статей, присвячених біологічним дослідженням і посиланням на повні тексти. Один із найважливіших таких сховищ – PubMed.

**PubMed** (<https://www.ncbi.nlm.nih.gov/pubmed>) – безкоштовна електронно-пошукова система, яка пропонує доступ до бази даних MEDLINE, посилань та резюме статей у журналах біомедицинської спрямованості.

PubMed містить: MEDLINE; PREMEDLINE; видавничі описи.

**MEDLINE** – база даних медичної інформації, яка містить бібліографічні описи з понад 4800 медичних періодичних видань зі всього світу, починаючи з початку 1960-х років. Нині MEDLINE доступна безкоштовно для пошуку через Інтернет як для фахівців, так і для широкої публіки.

MEDLINE складається з описів статей із медичних журналів і інших періодичних видань на 30 мовах (заголовки статей перекладаються англійською мовою). Приблизно 76 % описів містять реферати. Близько 9000 описів вносяться в MEDLINE щотижня – понад 571 000 на рік. Теми в MEDLINE мають широкий спектр галузей, які стосуються біології та медицини: наукові дослідження і їхні методології, клінічна практика, медсестринська справа, стоматологія, фармакологія, ветеринарія, а також суміжні дисципліни, зокрема медичні аспекти біології, зоології, ботаніки.

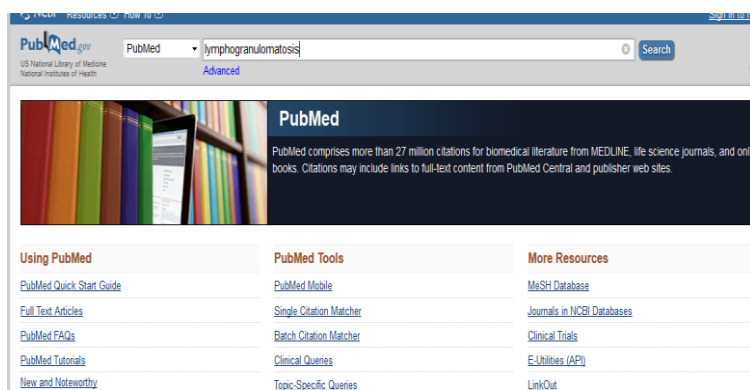
Детальний опис MEDLINE англійською мовою можна знайти в інформаційному листі NLM за адресою: <http://www.nlm.nih.gov/pubs/factsheets/medline.html>.

PREMEDLINE – файл бібліографічних описів, уведених у базу даних порівняно нещодавно і який ще не пройшов процедури індексування. У процесі індексування до кожного опису додається список наочних рубрик, які відображають змістовий вміст документа, та із спеціального пошукового словника – тезауруса, який називається MESH List (Medical Subject Headings List). Надалі пошук в PubMed може вестися за наочними рубриками, що підвищує повноту та точність пошуку.

Бібліографічні описи, які належать PREMEDLINE, мають спеціальну позначку [PubMed – in process]. Після індексування документи з PREMEDLINE переміщуються в MEDLINE. Файл видавничих описів містить записи, отримані від видавництва електронним методом. Вони позначені позначкою [PubMed – as supplied by publishers]. Видавничі описи не мають при собі наочних рубрик.

Бібліографічні описи, які не пройшли процедури індексування, не будуть знайдені пошуком у наукових рубриках. Ці описи потрібно шукати іншими способами – за текстом, заголовками або вихідними даними.

Головний пошуковий екран PubMed показаний на рис. 13.



*Рис. 13. Головний пошуковий екран PubMed*

Нагорі екрана розміщене пошукове віконце, в якому потрібно надрукувати англійською інформаційний запит, наприклад медичний термін, прізвище автора або заголовок журналу.

---

PubMed дає змогу здійснювати пошук різного ступеня складності. Простим є пошук за ключовими словами, тобто термінами, які виражають основний зміст інформаційного запиту.

Самі статті у PubMed належать журналам, у яких вони опубліковані, і доступ до повного тексту більшості з них не безкоштовний. PubMed пропонує доступ до деяких журналів з інших галузей, повний текст деяких статей та дуже зручну систему пошуку за посиланнями в обох напрямках (тобто на що посилається кожна стаття та які статті посилаються на неї).

**Bookshelf** (<http://www.ncbi.nlm.nih.gov/books>) – це оперативний ресурс, який забезпечує вільний доступ до повного тексту онлайн версії популярних підручників із біологічних наук (всі англійською мовою). Створений і обслуговується NCBI і національною медичною бібліотекою США (National Library of Medicine (NLM)). Bookshelf доповнює інші біомедичні літературні ресурси в NCBI (наприклад PubMed).

Ресурс створено в 1999 як енциклопедичне посилання для ресурсів PubMed і GenBank. Сьогодні ресурс нараховує понад 1500 тем. На відміну від інших баз даних NCBI, наприклад GenBank і Gene, які мають суворо визначену структуру даних, у Bookshelf книги входять у різних форматах, типах, розмірах і авторських розробках.

Книги, які є у базі даних, охоплюють матеріал із біохімії, молекулярної біології, клітинної біології, генетики, медицини, методів біологічних досліджень та вірусології.

**Scopus** (<https://www.scopus.com>) – бібліографічна і реферативна база даних і інструмент для відстежування цитованості статей, опублікованих у наукових виданнях.

База даних Scopus позиціонується видавничою корпорацією Elsevier як найбільша у світі універсальна реферативна база даних з можливостями відстеження наукової цитованості публікацій. За оголошеною стратегією ця база даних має стати найповнішим та вичерпним ресурсом для пошуку наукової літератури. Класифікаційна система SciVerse Scopus охоплює 24 тематичні розділи.

Scopus індексує 18 тис. назв наукових видань із технічних, медичних і гуманітарних наук. База даних індексує наукові журнали, матеріали конференцій і серійні книжкові видання.

---

---

Програмне забезпечення баз даних повинно задовольняти такі функціональні вимоги:

1. Необхідно, щоб обсяг баз даних був майже необмеженим (обмежений лише параметрами апаратних засобів).

2. База даних повинна бути достатньо гнучкою для забезпечення процесу перебудови в міру її заповнення, оскільки попереднє проектування детальної структури бази даних неможливе.

3. Бази даних повинні бути інтегровані з різними міжнародними базами даних, постійно оновлюватися, синхронізуватися з аналогічними базами даних та мати системи перевірки даних, які надходять.

4. Бази даних мають підтримувати не лише стандартні мультимедійні формати, але й багато спеціальних гіпермедіа-середовищ (просторові структури молекул, хімічні структурні формули, метаболічні шляхи).

5. Експлуатація та поповнення баз даних повинні бути легкодоступними та зрозумілими для користувачів без комп'ютерної підготовки (біологи, медики).

Усі наявні бази даних надають можливість роботи з ними через Інтернет та майже всі вони використовують стандартні методику пошуку.

Робота з базами даними може відбуватися в двох режимах.

1. Offline-інтерфейс – спочатку з мережі Інтернет на локальний комп'ютер скачується частина бази даних, потім із цією частиною виконується подальша робота.

2. Режим клієнт-сервер – на локальному комп'ютері встановлюється програма математичної обробки нуклеотидних або амінокислотних послідовностей, далі дана програма з'єднується з сервером бази даних і обробляє інформацію без скачування на локальний комп'ютер.

Водночас, інтенсивно розвиваються системи обробки інформації та пошукові системи, які збирають і обробляють інформацію відповідно до запитів користувачів.

На сьогодні в мережі Інтернет існує близько тисячі різних баз даних, але найпоширеніші ті, які містять первинну біологічну інформацію.

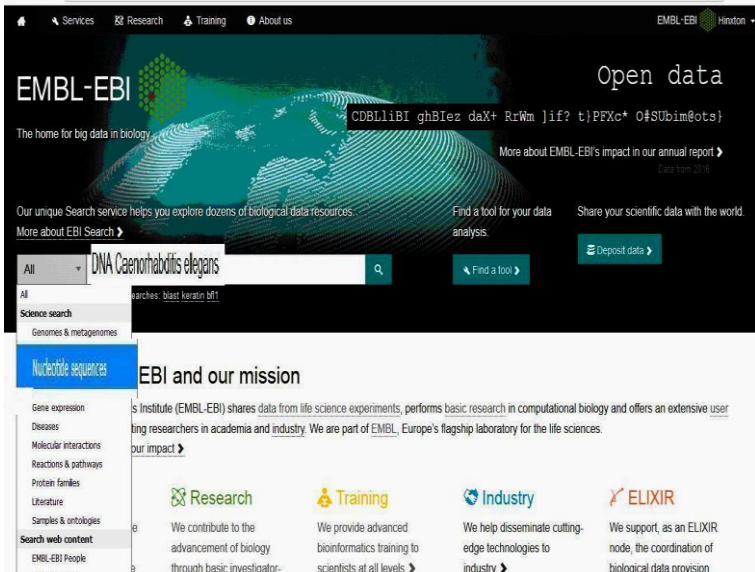
## ПРАКТИКУМ

### Завдання 1

Знайдіть у банку EMBL усі записи, які містять інформацію про ДНК нематод (*Caenorhabditis elegans*). Порівняйте ці результати з результатами з бази даних GenBank.

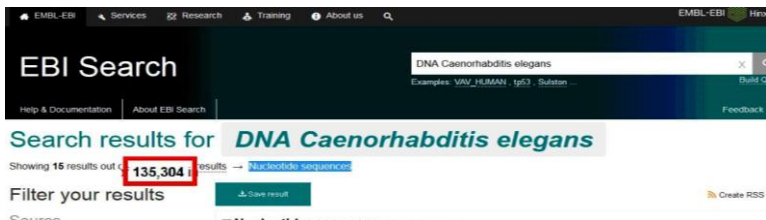
### Рекомендація

1. Здійснюємо вхід на веб-ресурс EMBL – <http://www.ebi.ac.uk>. У пошуковому полі вводимо запит для ДНК нематоди (DNA *Caenorhabditis elegans*), у лівому меню від пошукового поля вибираємо «Nucleotide sequences» і натискаємо «Search».



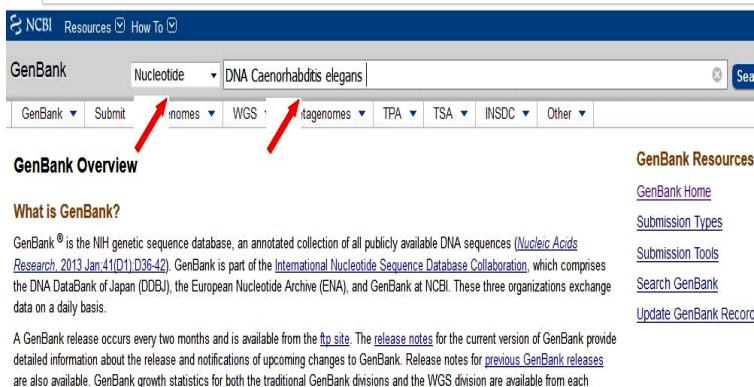
The screenshot shows the EMBL-EBI homepage. At the top, there are navigation links for Services, Research, Training, and About us. The main header features the EMBL-EBI logo and the text 'Open data'. Below this, there is a search bar with the query 'DNA Caenorhabditis elegans'. The left sidebar menu is open, and 'Nucleotide sequences' is selected. The main content area displays 'EBI and our mission' with a brief description of the institute's role and four icons representing Research, Training, Industry, and ELIXIR.

2. Із результатів пошуку випливає, що кількість записів у бази даних EMBL становить 135304.



The screenshot shows the EBI Search results page. The search bar contains 'DNA Caenorhabditis elegans'. Below the search bar, it says 'Showing 16 results out of 135,304 results'. The number '135,304' is highlighted with a red box. The left sidebar shows 'Filter your results' and 'Save result' buttons. The main content area shows the search results for 'DNA Caenorhabditis elegans'.

3. Для порівняння з іншою базою даних нуклеїнових кислот здійснюємо вхід на веб-ресурс GenBank (<http://www.ncbi.nlm.nih.gov/Genbank>). У пошуковому полі вводимо запит для ДНК нематоди зліва «Nucleotide» і натискаємо «Search».



**GenBank Overview**

**What is GenBank?**

GenBank<sup>®</sup> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan 41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each

4. Із результатів пошуку випливає, що кількість записів у базі даних GenBank становить 15951.



**Nucleotide** Nucleotide DNA Caenorhabditis elegans Search

Species: Animals (5,389), Plants (575), Fungi (2,022), Protists (554), Bacteria (7,299), Archaea (63), Viruses (18), Customize...

Molecule types: genomic DNARNA (13,494), mRNA (2,413), Customize...

Source databases: INSDC (GenBank) (13,916)

Summary: 20 per page Sort by Default order Send to: Filters: [Manage Filters](#)

See [etf5 \(DNA\) chromomethylase](#) in the Gene database  
[dna](#) reference sequences [Transcript \(1\)](#) [Protein \(1\)](#)

Items: 1 to 20 of **15951**

Found 24473 nucleotide sequences. Nucleotide (15951) EST (8464) GSS (58)

[Photorhabdus asymbiotica ATCC43949 complete genome](#)

1. 5,064,808 bp circular DNA  
Accession: FM162591.1 GI: 253778933  
[Assembly](#) [BioProject](#) [BioSample](#) [Protein](#) [Published](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

Results by taxon

Top Organisms: [Tree](#)  
Caenorhabditis elegans (3463)  
Clostridium difficile (1732)  
Vibrio cholerae (594)  
Vibrio parahaemolyticus (568)  
Brucella abortus (188)  
All other taxa (9406)  
More...

Find related data  
Database: Select

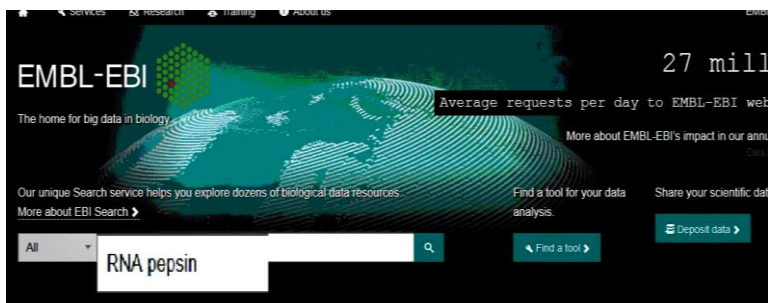
## Завдання 2

Скільки записів міститься у базі даних EMBL-EBI про РНК пепсину? Яка кількість цих записів стосується лише організму людини?

## Рекомендація

1. Здійснюємо вхід на веб-ресурс EMBL (<http://www.ebi.ac.uk>).

У пошуковому полі вводимо запит для РНК пепсину (RNA pepsin) і натискаємо «Search».



### Explore EMBL-EBI and our mission

The European Bioinformatics Institute (EMBL-EBI) shares data from life science experiments, performs basic research in computational biology and offers an extensive training programme, supporting researchers in academia and industry. We are part of EMBL, Europe's flagship laboratory for the life sciences.

2. Результати пошуку засвідчують, що кількість записів у базі даних EMBL становить 952.



3. Для встановлення кількості записів для РНК пепсину людини у пошуковому полі вводимо «human RNA pepsin» і натискаємо «Search». Кількість записів становить 362.



### Завдання 3

Знайдіть у базі даних NDB усі записи, які містять інформацію про: А-форму молекули ДНК, комплекси ДНК із ліками та білками.

#### Рекомендація

1. Здійснюємо вхід на веб-ресурс NDB (<http://ndbserver.rutgers.edu>). У меню головного пошукового екрана вибираємо блок «Search DNA».

2. Переходимо на сторінку, яка містить опції цього блоку. В меню зліва вибираємо «A DNA». Отримуємо результат. Кількість записів про А-форму молекули ДНК становить 414.

The screenshot shows the NDB Nucleic Acid Database search results for the query "Drug DNA Complexes". The search options are set to "Polymer: Drug DNA Complexes". The results show 414 entries. The first entry is titled "Structure of 8-mer DNA2" with NDB ID: 5XX0 and PDB ID: 5XX0. The second entry is titled "Structure of 8-mer DNA3" with NDB ID: 5XX1 and PDB ID: 5XX1. The third entry is titled "RNA-DNA heptamer duplex with one 2'-5'-linkage" with NDB ID: 5U0Q and PDB ID: 5U0Q.

3. Для встановлення кількості записів щодо комплексів ДНК із ліками у меню блоку зліва вибираємо «Drug DNA Complexes». Із результатів пошуку видно, що кількість записів, які стосуються комплексів ДНК із ліками в базі даних NDB становить 401.

The screenshot shows the NDB Nucleic Acid Database search results for the query "Protein DNA Complexes". The search options are set to "Polymer: Protein-DNA-Complexes". The results show 401 entries. The first entry is titled "Light-activated ruthenium complex bound to a DNA quadruplex" with NDB ID: 5L58 and PDB ID: 5L58. The authors listed are McQuaid, K.T., Abell, H., Brazier, J., Cardin, C.J., Hall, J.P. The citation is "Chiral recognition of a DNA quadruplex by a light-activated ruthenium complex To Be Published pp. - 0". The experiment is "X-RAY DIFFRACTION Resolution:1.78Å R work:0.214 R free:0.232".

4. Щоб установити кількість записів щодо комплексів ДНК із білками, в меню блоку зліва вибираємо «Protein DNA Complexes». Із результатів пошуку видно, що кількість записів, які стосуються комплексів ДНК з білками, в базі даних NDB становить 4375.

### Завдання 4

Знайдіть у базі даних UniProt усі записи білків представників родини пасльонові (*Solanaceae*).

### Рекомендація

1. Здійснюємо вхід на веб-ресурс UniProt (<http://www.uniprot.org>). Вводимо запит у пошуковому полі UniProt для родини пасльонові (*Solanaceae*) і натискаємо «Search».

2. Результати пошуку показують, що кількість записів білків представників родини пасльонові у базі даних UniProt становить 336396.

UniProtKB results

Filter by:

Entry	Entry name	Protein names	Gene names	Organism
P00876	RBL_TOBAC	Ribulose biphosphate carboxylase L...	rblct, rbcA	Nicotiana tabacum (Common tobacco)
P05117	PGLR_SOLLC	Polygalacturonase-2	PG2 PG, PG2A, PG2B	Solanum lycopersicum (Tomato) (Lycopersicon esculenta)
ERL745	PAT_PETHY	Bifunctional aspartate aminotransfe...		Petunia hybrida (Petunia)
Q40392	TMVRM_NCCUJ	TMV resistance protein N	N	Nicotiana glauca (Tobacco)
P80042	GGPS_CAPAN	Geranylgeranyl pyrophosphate syntha...	GGPS1	Capicum annuum (Bell pepper)

### Завдання 5

Використовуючи базу даних UniProt встановіть: до якого класу ферментів належить цитозольна фосфоліпаза  $A_2$  миші? Які ліганди може приєднувати фосфоліпаза  $A_2$ ?

#### Рекомендація

1. Здійснюємо вхід на веб-ресурс UniProt (<http://www.uniprot.org>). У полі пошуку вводимо ідентифікатор UniProt для фосфоліпазу  $A_2$  миші (*mus musculus phospholipase A<sub>2</sub>*) і натискаємо «Search».

UniProtKB

mus musculus phospholipase A2

UniProtKB

UniProt Knowledgebase

Swiss-Prot (556,388)

Manually annotated and reviewed.

Records with information extracted from literature and

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

News

Forthcoming changes  
Planned changes for UniProt

UniProt release 2017\_12  
Swiss-Prot in the sky with psilocybin: the biosynthesis pathway a psychedelic drug unravelled

2. Із результатів пошуку вибираємо той, який відповідає фосфоліпазі  $A_2$  миші (наприклад з ідентифікатором P47713).

UniProtKB results for **mus musculus phospholipase a2**

Filter by: Reviewed (54), Unreviewed (75), Popular organisms (Mouse, MmV, MmM, TmM, TmM, TmM), Search terms (gene ontology, keyword, protein family).

Entry	Entry name	Protein names	Close names
P10107	ANKA1_MOUSE	Annexin A1	Anxa1 Anx1, Lpc-1, Lpc1
P07356	ANKA2_MOUSE	Annexin A2	Anxa2 Anx2, Cal1h
Q8R3U1	PA216_MOUSE	HRAS-like suppressor 3	Pla2g16 H-rev107, Hrasl3, Hrev107
Q62026	PLA2R_MOUSE	Secretory phospholipase A2 receptor	Pla2r1 Pla2g1br
P47713	PA24A_MOUSE	Cytosolic phospholipase A2	Pla2g4a Cpla2, Pla2g4
P97819	PLPL9_MOUSE	85/88 kDa calcium-independent phosph...	Pla2g6 Ppla9
Q08700	PRDX6_MOUSE	Peroxiredoxin-6	Prdx6 Aop2, Ltx4, Prdx5

3. Переходимо на сторінку результату (P47713) та опрацюємо потрібну інформацію. Встановлено, що фосфорліпаза A<sub>2</sub> миші належить до третього класу ферментів, оскільки EC:3.1.1.4. Лігандами ферменту є метали, зокрема кальцій.

View the complete GO annotation on QuickGO ...

**Keywords**<sup>1</sup>

Molecular function	Hydrolase
Biological process	Lipid degradation, Lipid metabolism
Enzyme class	Ligand, Calcium, Metal-binding

BRENDA<sup>4</sup> 3.1.1.4, 3474.  
 Reactome<sup>2</sup> R-MMU-111995, phospho-PLA2 pathway, R-MMU-1482788, Acyl chain remodeling of PC, R-MMU-1482798, Acyl chain remodeling of CL, R-MMU-1482801, Acyl chain remodeling of PS, R-MMU-1482839, Acyl chain remodeling of PE, R-MMU-1482922, Acyl chain remodeling of PI, R-MMU-1482925, Acyl chain remodeling of PG, R-MMU-1483115, Hydrolysis of LPC, R-MMU-1483166, Synthesis of PA, R-MMU-2142753, Arachidonic acid metabolism, R-MMU-418592, ADP signalling through P2Y purin, R-MMU-432142, Platelet sensitization by LDL, R-MMU-6811436, COPI-independent Golgi-to-ER f

SABIO-RK<sup>3</sup> P47713.

**Names & Taxonomy**<sup>1</sup>

Protein names<sup>4</sup>

**Recommended name:**  
**Cytosolic phospholipase A2**  
 ■ Short name: cPLA2  
 Alternative name(s):  
 ■ Phospholipase A2 group IVA

◆ Phospholipase A2 (EC:3.1.1.4)

### Завдання 6

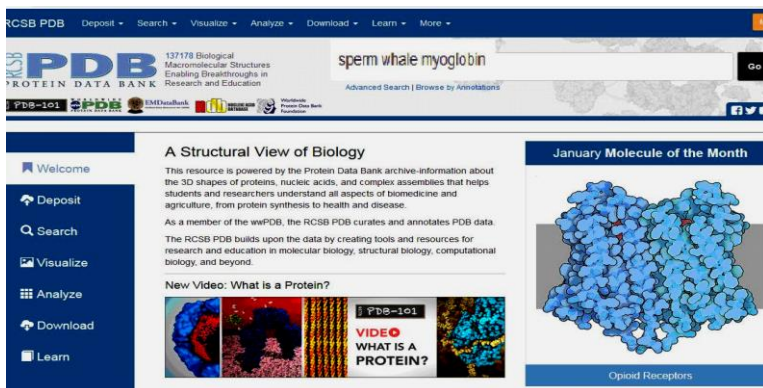
Знайдіть третинну структуру міоглобіну кита (sperm whale myoglobin) у Protein Data Bank. Вкажіть його ID.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс PDB (<https://www.rcsb.org>). У полі пошуку головного екрана PDB вводимо запит для



міоглобіну кита (sperm whale myoglobin) і натискаємо «Go».



2. Із результатів пошуку знаходимо міоглобін кита з ID 1VXA

280 Structures 27 Unreleased Structures 88 Citations 66 Ligands 3 News & PDB-101 Articles

Search Parameter:  [Refine Search](#) [Show Search to MyPDB](#)

Currently showing 1 - 25 of 280 Page: 1 of 12 [Previous](#) [Next](#) → Displaying 25 Results

**Refinements**

ORGANISM

- Physeter catodon (270)
- Aplysia limacina (5)
- Phacodes pectinatus (2)
- Phoca vitulina (1)
- Kogia sima (1)

UNIPROT MOLECULE NAME

- Myoglobin (272)
- Globin (6)
- Hemoglobin-1 (2)
- [Refine Query](#)

**1VXA**

**NATIVE SPERM WHALE MYOGLOBIN**

**Yang, F., Phillips Jr., G.N.**  
(1996) J Mol Biol **256** 762-774

**Released:** 8/1/1996  
**Method:** X-ray Diffraction  
**Resolution:** 2.0 Å

**Macromolecule:**  
MYOGLOBIN (protein)  
**Unique Ligands:** HEM, SO4

[Download File](#) [View File](#)

3. Переходимо на сторінку результату. Отримуємо 3D-структуру міоглобіну кита.

CSB PDB Deposit Search Visualize Analyze Download Learn More

**PDB** 137178 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment

Biological Assembly 1

**1VXA**

NATIVE SPERM WHALE MYOGLOBIN

DOI: 10.2210/pdb1VXA/pdb

Classification: [OXYGEN STORAGE](#)

Organism(s): [Physaster catodon](#)

Deposited: 1996-01-09 Released: 1996-08-01

Deposition Author(s): [Yang, F., Phillips Jr., G.M.](#)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2 Å

R-value Work: 0.177

wwPDB Validation

Metric	Percentile Ranks	Value
Cis-peptide	28	28
Ramachandran outliers	2.0%	2.0%
Sidechain outliers	20.0%	20.0%
RSRZ outliers	1.3%	1.3%

3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers: Protein Workshop | Ligand Explorer

This is version 1.2 of the entry. See complete history.

## Завдання 7

Установіть, для яких ферментів субстратом є сахароза. Представники якого класу ферментів не виявляються у цьому переліку? Скільки ферментів проявляють активність при pH 3?

### Рекомендація.

1. Здійснюємо вхід на веб-ресурс BRENDA (<http://brenda-enzymes.info>). Унизу (під полем пошуку) в меню вибираємо опцію «Substrate». У полі пошуку головного екрана BRENDA вводимо запит для сахарози (sucrose) і натискаємо «start search».

**BRENDA**  
The Comprehensive Enzyme Information System

sucrose

Substrate contains

add search field delete search field **start search**

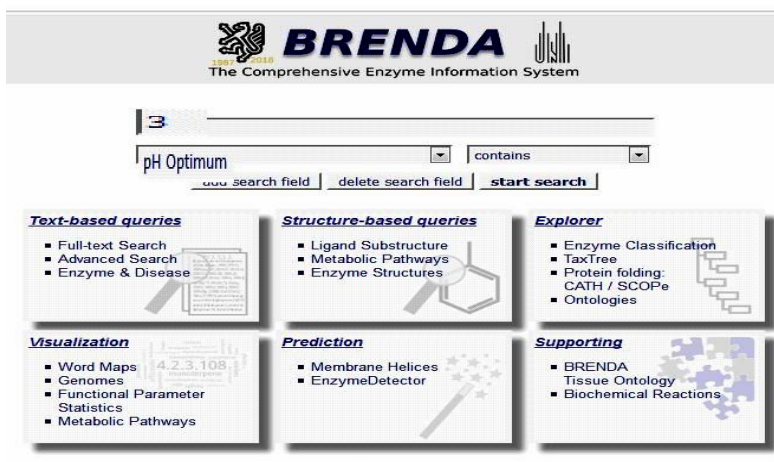
<p><b>Text-based queries</b></p> <ul style="list-style-type: none"> <li>Full-text Search</li> <li>Advanced Search</li> <li>Enzyme &amp; Disease</li> </ul>	<p><b>Structure-based queries</b></p> <ul style="list-style-type: none"> <li>Ligand Substructure</li> <li>Metabolic Pathways</li> <li>Enzyme Structures</li> </ul>	<p><b>Explorer</b></p> <ul style="list-style-type: none"> <li>Enzyme Classification</li> <li>TaxTree</li> <li>Protein folding: CATH / SCOPe</li> <li>Ontologies</li> </ul>
<p><b>Visualization</b></p> <ul style="list-style-type: none"> <li>Word Maps</li> <li>Genomes</li> <li>Functional Parameter Statistics</li> <li>Metabolic Pathways</li> </ul>	<p><b>Prediction</b></p> <ul style="list-style-type: none"> <li>Membrane Helices</li> <li>EnzymeDetector</li> </ul>	<p><b>Supporting</b></p> <ul style="list-style-type: none"> <li>BRENDA Tissue Ontology</li> <li>Biochemical Reactions</li> </ul>

2. Отримуємо результат пошуку, у якому вказані ферменти, для яких сахароза є субстратом. Таких ферментів виявлено 86. Проаналізувавши результати, внаслідок пошуку не виявлено ферментів класу 6.



EC Number ▼▲	Recommended Name ▼▲	Substrate ▼▲
1.1.1.47	glucose 1-dehydrogenase [NAD(P)+]	sucrose
1.1.3.9	galactose oxidase	sucrose
1.1.3.10	pyranose oxidase	sucrose
1.1.99.13	glucoside 3-dehydrogenase	sucrose
1.1.99.29	pyranose dehydrogenase (acceptor)	sucrose
2.3.1.79	maltose O-acetyltransferase	sucrose
2.4.1.B1	loliose synthase	sucrose
2.4.1.2	dextrin dextranase	sucrose
2.4.1.2	dextrin dextranase	xylosucrose
2.4.1.4	amylsucrase	sucrose

3. Для визначення кількості ферментів, які проявляють свою активність при pH 3, унизу (під полем пошуку) в меню вибираємо опцію «pH Optimum». У полі пошуку головного екрана BRENDA вводимо значення 3 і натискаємо «start search».



4. Отримуємо результат пошуку, в якому вказано 89 ферментів, які проявляють свою активність при pH=3.























Advanced Search

## Search pH Optimum

Search term: 3

« Results 1 - 10 of 89 »  
 download as CSV  
 download all results as CSV

EC Number ▲	Recommended Name ▼	pH Optimum Minimum ▼
  1.1.1.1	alcohol dehydrogenase	3
  1.1.1.71	alcohol dehydrogenase [NAD(P)+]	3
  1.1.3.7	aryl-alcohol oxidase	3
  1.1.3.13	alcohol oxidase	3
  1.1.5.2	glucose 1-dehydrogenase (PQQ, quinone)	3
  1.1.99.18	cellobiose dehydrogenase (acceptor)	3
  1.2.3.1	aldehyde oxidase	3
  1.2.3.4	oxalate oxidase	3
  1.8.2.2	thiosulfate dehydrogenase	3
  1.10.3.2	laccase	3

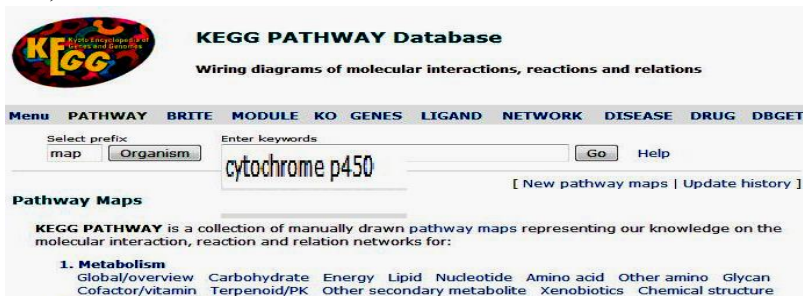
« Results 1 - 10 of 89 »  
 download as CSV

## Завдання 8

Зобразить метаболічний шлях для цитохрому P450.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс KEGG PATHWAY ([www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)). Якщо спочатку зайшли на головну сторінку сервера KEGG, то потрібно вибрати блок KEGG PATHWAY. У полі пошуку головного екрана KEGG PATHWAY вводимо запит для цитохрому P450 (cytochrome P450). Натискаємо «Go».



**KEGG PATHWAY Database**  
 Wiring diagrams of molecular interactions, reactions and relations

Menu **PATHWAY** BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

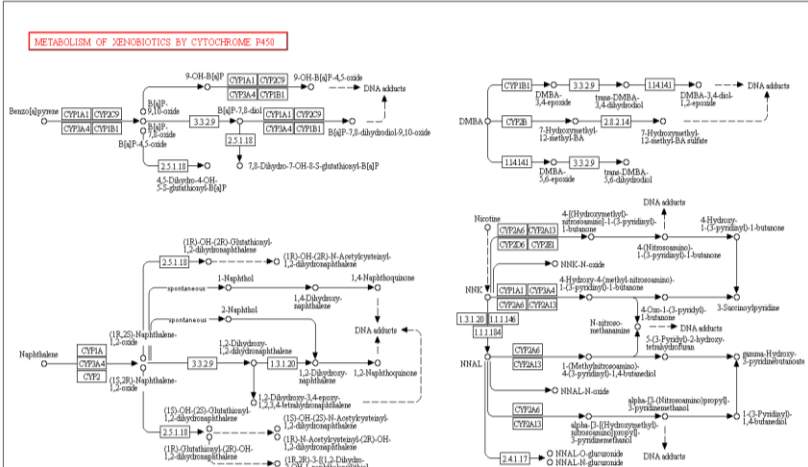
Select prefix: map  Enter keywords:

**Pathway Maps** [ New pathway maps | Update history ]

**KEGG PATHWAY** is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for:

**1. Metabolism**  
 Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan  
 Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure

2. Отримуємо результати метаболічних шляхів. Здійснюємо вхід на один із них і аналізуємо його.



## Завдання 9

Використовуючи базу даних Таксоному, вкажіть скільки класів містить відділ Червоних водоростей (Rhodophyta). Чи наявна дана інформація у базі даних Species 2000?

## Рекомендація

1. Здійснюємо вхід на веб-ресурс Таксоному (<http://www.ncbi.nlm.nih.gov/Taxonomy>). У полі головного пошукового екрана вводимо запит для відділу Червоних водоростей (Rhodophyta) і натискаємо «Search».

The screenshot shows the NCBI Taxonomy database search results for Rhodophyta. The search bar contains "Rhodophyta" and the results show a list of taxonomic levels. Below the search bar, there is a grid of butterfly images and a section titled "Taxonomy" with the text: "The Taxonomy Database is a curated classification and nomenclature for all of the organism databases. This currently represents about 10% of the described species of life on the plan".

Using Taxonomy

Taxonomy Tools

Other Resources

2. Отримуємо результат та підраховуємо кількість класів.

NCBI Taxonomy Browser

Search for: as complete name

Display: 3 levels using filter: none

- [Rhodophyta \(red algae\)](#)
  - [Bangiophyceae](#)
    - [Bangiales](#)
      - [Bangiaceae](#)
      - [unclassified Bangiales](#)
    - [Cyanidiales](#)
      - [Cyanidiaceae](#)
      - [unclassified Cyanidiales](#)
      - [environmental samples](#)
    - [Porphyridiales](#)
      - [Phragmonemataceae](#)
      - [Porphyridiaceae](#)
    - [unclassified Bangiophyceae](#)
      - [Chroothece richteriana](#)
    - [environmental samples](#)
      - [uncultured Bangiophyceae](#)
  - [Compsopogonophyceae](#)
    - [Compsopogonales](#)
      - [Baldiaceae](#)
      - [Compsopogonaceae](#)
    - [Erythropeltidales](#)
      - [Erythropeltidiaceae](#)
      - [Erythrotrichiaceae](#)
      - [unclassified Erythropeltidales](#)
    - [Rhodochaetales](#)
      - [Rhodochaetaceae](#)
      - [environmental samples](#)

3. Здійснюємо вхід на веб-ресурс Species 2000 (<http://www.sp2000.org>). У полі головного пошукового екрана вводимо запит для відділу Червоних водоростей (Rhodophyta) і натискаємо «Search».

Species 2000

Search: Rhodophyta

Home

About

Board of Directors

3. Із отриманих результатів видно, що інформації про відділ Червоних водоростей у базі даних Species 2000 немає.

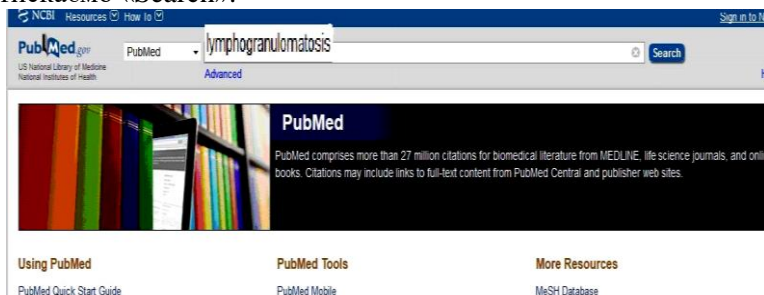


## Завдання 10

Скільки безкоштовних статей про лімфогрануломатоз міститься у PubMed? Скільки статей опубліковано у 2017 році?

### Рекомендація

1. Здійснюємо вхід на веб-ресурс PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>). У полі пошуку вводимо запит для лімфогрануломатозу (lymphogranulomatosis) і натискаємо «Search».



2. Із результатів пошуку видно, що кількість статей у базі даних PubMed із лімфогранулематозу – 72232.



3. Щоб установити кількість безкоштовних статей у лівому меню вибираємо опцію «Free full text».

Із результатів видно, що число безкоштовних статей – 13514.



The screenshot shows the PubMed search results page for the query "lymphogranulomatosis". The search results are filtered to show only "Free full text" articles, resulting in 13,514 items. The first result is titled "Hospitalization Rates, Prevalence of Cardiovascular Manifestations, and Outcomes Associated With Sarcoidosis in the United States" by Patel N, Kalaria R, Doshi R, Arora H, Bajaj NS, Arora G, Arora P. The article is published in the J Am Heart Assoc in 2018. The search results page also includes a "Results by year" histogram on the right side, which is partially visible.

4. Для встановлення кількості опублікованих статей у 2017 році аналізуємо гістограму, яка є праворуч на сторінці результату пошуку веб-ресурсу PubMed. Встановлено, що кількість статей за 2017 рік рівний 529.



The screenshot shows the PubMed search results page for the query "lymphogranulomatosis". The search results are filtered to show only "Free full text" articles, resulting in 13,514 items. The first result is titled "Hospitalization Rates, Prevalence of Cardiovascular Manifestations, and Outcomes Associated With Sarcoidosis in the United States" by Patel N, Kalaria R, Doshi R, Arora H, Bajaj NS, Arora G, Arora P. The article is published in the J Am Heart Assoc in 2018. The search results page also includes a "Results by year" histogram on the right side. A callout box highlights the number of articles published in 2017, which is 529.

---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

### 1. Що є структурною одиницею банку даних UniProt?

- а) нуклеотидна послідовність;
- б) амінокислотна послідовність;
- в) просторові атомні координати;
- г) опис ферментативної реакції;
- д) структура метаболічного шляху.

2. Доповніть відповідь: *Залежно від організації інформації бази даних поділяються на:* а) .....; б) .....; в) .....

### 3. Що є структурною одиницею банку даних KEGG PATHWAY?

- а) нуклеотидна послідовність;
- б) амінокислотна послідовність;
- в) просторові атомні координати;
- д) структура метаболічного шляху;
- г) опис ферментативної реакції.

### 4. У яких базах даних можна знайти інформацію про білкові фолди?

- а) UniProt;
- б) SCOP;
- в) EMBL;
- г) DDBJ;
- д) HIV-SD.

### 5. Виберіть із переліку бази даних, які містять ідентичну інформацію:

- а) UniProt;
- б) SCOP;
- в) EMBL;
- г) GenBank;
- д) DDBJ;
- е) PDB.

6. Категорії бази даних SCOP розміщуються в порядку від найменшої до найбільшої:

- а) фолд → суперродина → родина → домен;
- б) домен → родина → суперродина → фолд;
- в) родина → суперродина → фолд → домен;
- г) домен → фолд → родина → суперродина.

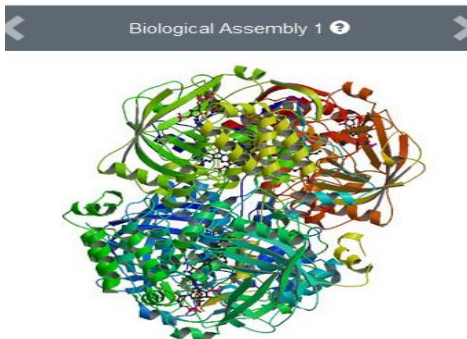
### 7. Що є структурною одиницею банку даних BRENDA?

- а) нуклеотидна послідовність;
- б) амінокислотна послідовність;

- в) просторові атомні координати;
- г) опис ферментативної реакції;
- д) структура метаболічного шляху.

**8. Структурним компонентом якої бази даних виявляється зображений фрагмент запису?**

- а) UniProt;
- б) SCOP;
- в) EMBL;
- г) CATH;
- д) PDB.



**9. Додаткові бази UniProt такі: .....**

- а) GenBank;
- б) EMBL;
- в) TrEMBL;
- г) CATH;
- д) PDB.

**10. Основними елементами якої бази даних є клас, архітектура, топологія, гомологія:**

- а) UniProt;
- б) REBASE;
- в) EMBL;
- г) CATH;
- д) GOBASE.

**11. Яка база даних містить інформацію про тривимірну структуру білків?**

- а) UniProt;
- б) ENZYME;
- в) EMBL;
- г) RDP;
- д) PDB.

## 12. Структурним компонентом якої бази даних виявляється зображений фрагмент запису?

P04040-1

[FASTA](#) [Add to basket](#)

```
10          20          30          40          50
MADSRDPASD QMQHWKEQRA AQAADVLTIG AGNPFVGDRLN VITVGRGRL
60          70          80          90         100
LVQDVVFDE MAHFDRERIF ERVVRHAKGAG AFGYFEVTHD ITKYSKRVF
110         120         130         140         150
EHIGKKTPIA VRFSTVAGES GSADTVRDDR GFAVKFYTED GNWDLVGNT
160         170         180         190         200
PIFFIRDPII FPFSEINSQKR NPQTHLKDED MVWDFWSLRP ESLHQVSELF
210         220         230         240         250
SDRGIPDGRH HMGYGSHTF KLVNANGEAV YCRPHYKTDQ GIRNLSVEDA
260         270         280         290         300
ARLSQEDDPY GIRDLENAIA TGKYPSTWTFY IQVMTEFNAE TFFENPFDLT
310         320         330         340         350
KVWPHKDYPL IPVVKLVLRN NPVNYFAEVE QIAFDPSNMP PGIEASPEKM
360         370         380         390         400
LQGRLEFAYD THRHRLGPNY LHIEVNCPIYR AEVANYQRDG PMCMQDNQGG
410         420         430         440         450
APNYEYNSIS APEQQPSALE HSIQYSGEVR RFNTANDDNV TQVRAFYVNV
460         470         480         490         500
LNEEQKRLC ENIAGHLKDA QIFIQKRAV NFEVHPDYG SHIQALLDKY
510         520
NAEKPRNAIH TFVQSGSHLA AREKANL
```

а) SCOP; б) EMBL; в) CATH; г) UniProt; д) PDB; е) KEGG.

## 13. Бази даних SCOP і CATH поповнюються за рахунок інформації, отриманої з бази даних .....

- а) Taxonomy;
- б) DDBJ;
- в) EMBL;
- г) BRENDA;
- д) PDB;
- е) GenBank.

## 14. Що є структурною одиницею банку даних EMBL?

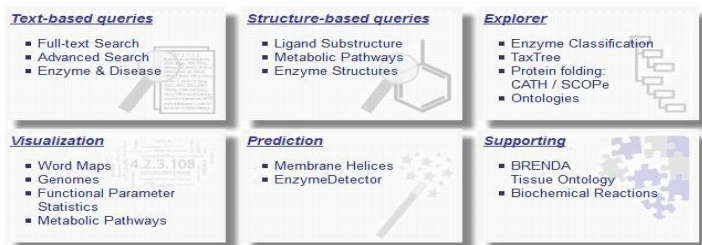
- а) нуклеотидна послідовність;
- б) амінокислотна послідовність;
- в) просторові атомні координати;
- г) опис ферментативної реакції;
- д) структура метаболічного шляху;

## 15. Які бази даних класифікують організми та можуть бути використані для визначення положення досліджуваного організму в ієрархії:

- а) UniProt;
- б) Taxonomy;
- в) DDBJ;
- г) Species 2000;
- д) BRENDA.



**16. У якій базі даних можна виявити таку інформацію .....**



а) SCOP; б) CATH; в) EMBL; г) BRENDA; д) PDB; е) GenBank?

**17. Із переліку виберіть бібліографічні бази даних .....**

- а) Scopus;
- б) REBASE;
- в) GOBASE;
- г) PubMed;
- д) NCBI Bookshelf.

**18. TrEMBL складається із двох основних секцій REM-TrEMBL і SP-TrEMBL. Яка із цих секцій містить послідовності, отримані синтетично, неповні і не цікаві для анотування?**

**19. Що є структурною одиницею банку даних PDB?**

- а) нуклеотидна послідовність;
- б) амінокислотна послідовність;
- в) просторові атомні координати білків;
- г) опис ферментативної реакції;
- д) структура метаболічного шляху;

**20. Яка база даних містить інформацію із ферментів і сайтів рестрикції?**

- а) UniProt;
- б) SCOP;
- в) EMBL;
- г) REBASE;
- д) PDB;
- е) GenBank.

**21. Структурним компонентом якої бази даних є зображений фрагмент запису?**



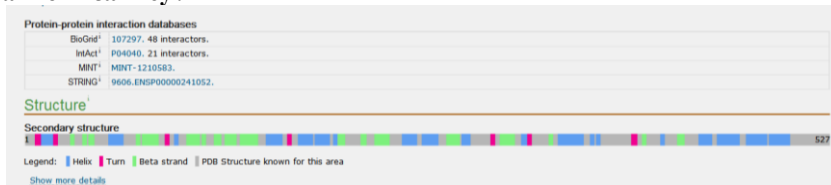
25. Структурним компонентом якої бази даних є зображений фрагмент запису?

```
ORIGIN
      1 atgaaggctc tcaccccttc ctgtctgggt gctctggcca ttgctcgagg gggttctcac
     61 catcaccatc accatccatg ggattacaag gacgacgatg acaagtttgt gaaccaaacac
    121 ctgtgccggt cacaccctgt ggaagctctc tacctagttg gcggggaacg aggcctcttc
    181 tacacaccca agaccgcgag ggaagcagag gacctgcagg tggggcaggt ggaagtgggc
    241 gggggccctg gtgcaggcag cctgcagccc ttggccctgg aggggtccct gcagaagcgt
    301 ggcattgttg aaacaatgctg taccagcacc tgctccctct accagctgga gaactactgc
    361 aactag
```

а) SCOP; б) GenBank; в) CATH; г) UniProt; д) PDB; е) KEGG.

26. База даних PIR поділена на чотири секції: PIR1, PIR2, PIR3 і PIR4. Яка з цих секцій повністю класифікована за суперродинами й анотована?

27. Структурним компонентом якої бази даних є зображений фрагмент запису?



а) UniProt; б) SCOP; в) EMBL; г) CATH; д) PDB.

28. Архівні бази даних, які містять анотовані первинні структури ДНК і білків, просторові структури нуклеїнових кислот і білків, а також профілі експресії генів білків клітин належать до .....

- а) первинних;
- б) вторинних.

29. Виберіть правильну відповідь: Єдина база даних UniProt створена об'єднанням баз даних:

- а) PIR;
- б) TrEMBL;
- в) EMBL;
- г) CATH;
- д) Swiss-Prot.

30. Виберіть правильну відповідь: За тематикою бази даних часто поділяють на два типи – бази даних загального та спеціального призначень. Бази даних ДНК, білків, вуглеводів є базами .....

- а) загального призначення;
- б) спеціального призначення.

---

---

## РОЗДІЛ 2 МЕТОДИ БІОІНФОРМАЦІЙНИХ ДОСЛІДЖЕНЬ

### ТЕМА 3. ВИРІВНЮВАННЯ ПОСЛІДОВНОСТЕЙ І ФІЛОГЕНЕТИЧНИЙ АНАЛІЗ

**Мета:** ознайомитися з теоретичними основами парного та множинного вирівнювання. Навчитися здійснювати різні типи вирівнювань для встановлення структурних, функціональних і еволюційних відносин між послідовностями біологічних макромолекул.

#### ТЕОРЕТИЧНА ЧАСТИНА

##### 1. Завдання та види вирівнювання послідовностей

Вирівнювання послідовностей дає змогу встановити відповідність залишків у ланцюгах, які є основним засобом у біоінформатиці. Методи порівняння послідовностей нуклеотидів нуклеїнових кислот та амінокислот білків допомагають знаходити функціонально подібні або еволюційно споріднені об'єкти.

*Вирівнювання послідовностей* у біоінформатиці – метод порівняння нуклеотидних (ДНК, РНК) або амінокислотних (білки) послідовностей знаходженням схожих ділянок, що може бути наслідком функціональних, структурних або еволюційних зв'язків між послідовностями.

Порівняння нових нуклеотидних послідовностей ДНК і РНК, а також амінокислотних послідовностей білків відбувається на основі уже відомих, розшифрованих послідовностей та необхідне для виявлення загальних структурних і функціональних особливостей, а також для дослідження еволюційних зв'язків.

Мета вирівнювання послідовностей полягає в тому, щоб визначити ступінь подібності двох послідовностей і якщо вона достатньо висока зробити правомірний висновок про їхню гомологічність.

Під час передачі генетичної інформації від попереднього покоління наступним вона може змінюватися під час процесу реплікації. Зміни, які відбуваються у процесі відхилення від загального предка можуть бути трьох типів: заміни, вставки і

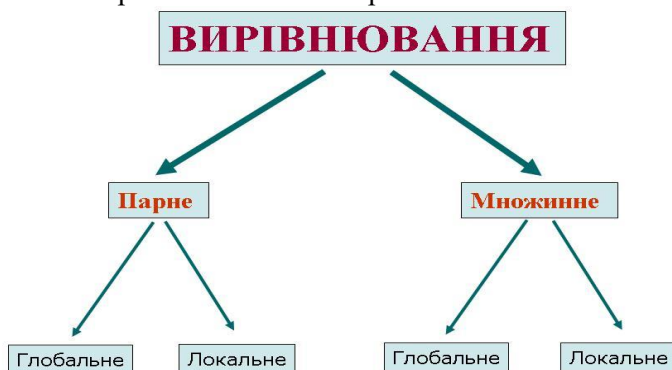
делеції. Ці зміни накопичуються від покоління до покоління, унаслідок чого може відбуватися значна кількість розбіжностей. Порівняння двох послідовностей показує ступінь їхнього відхилення.

Маючи дві або більше послідовностей дослідник може виконати такі *завдання*:

- визначити їхню спорідненість;
- розкрити відповідність залишків (мономерів);
- встановити консервативні та варіабельні ділянки;
- окреслити еволюційні взаємозв'язки.

Вирішити ці завдання можливо з використанням інформації, яка зберігається у базах даних. Значною мірою саме для виконання таких завдань і створювалися бази даних.

Залежно від кількості послідовностей, які порівнюються, розрізняють парне та множинне вирівнювання.



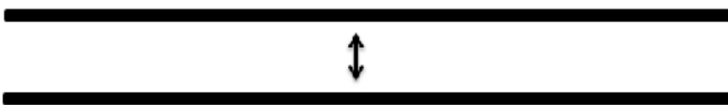
**Рис. 14. Види вирівнювання послідовностей нуклеотидів та амінокислот**

Парне вирівнювання здійснюється між двома послідовностями. Вирівнювання більш ніж двох послідовностей називається множинним.

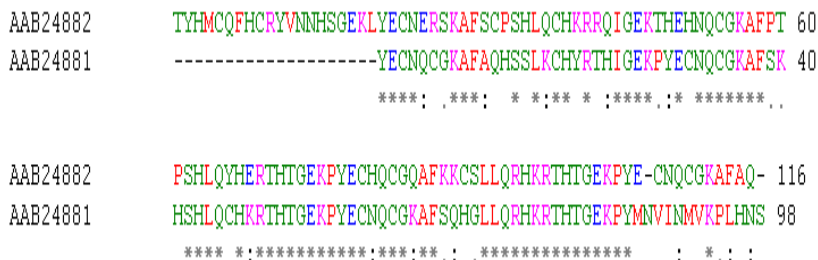
Незалежно від того, чи проводимо ми парне або множинне вирівнювання, в кожному з варіантів можна використовувати як глобальне, так і локальне вирівнювання.

1. Глобальне вирівнювання – форма глобальної оптимізації, яка знаходить відповідності нуклеотидних або амінокислотних

залишків по всій довжині досліджуваних послідовностей. Глобальне вирівнювання може бути застосоване, якщо дві послідовності гомологічні по всій довжині.



Результати вирівнювання послідовностей розміщуються в рядках матриці так, що елементи, які збігаються (нуклеотиди або амінокислоти), розміщені один під одним (в одній колонці).



**Рис. 15. Глобальне вирівнювання послідовностей амінокислот**

У даному типі вирівнювання символом «\*» позначені відповідності (однакові); «.» – подібні за властивостями; «>» – близькі за властивостями; «прогалини» позначають невідповідності, «-» позначає вставки (інсерції, від англ. insertion) і видалення (делеції, від англ. deletion), які необхідно зробити в обох послідовностях, щоб досягнути максимальної кількості відповідностей. Слід зауважити, що делеція в одній послідовності однозначна зі вставкою в іншій.

```

Метод. Первинної. -----Культури. Клітин.
| | | | |      | | | | |      | | | | | | | | | | |
Метод. Вторинної. Гетероплоїдної. Культури. Клітин.
    
```

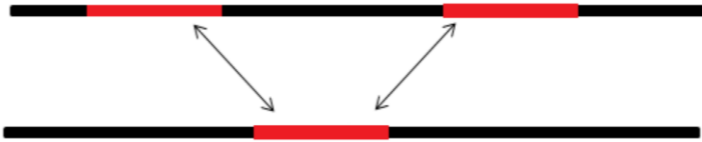
2. Локальне вирівнювання ідентифікує подібні ділянки у межах довгих послідовностей, які суттєво відрізняються на

---

---

більшій частині своєї протяжності.

Локальне вирівнювання – може бути застосоване для порівняння послідовностей із частковою гомологією.



Для локального збігу виступаючі кінці не розглядаються як пропуски (делеції). Можливі також вставки та видалення всередині збігаючої частини.

Первинної. Культури. Клітин.  
| | | | | | | | | | | | | | | |  
Метод. Вторинної. Гетероплоїдної. Культури. Клітин.

Для глобального вирівнювання застосовують алгоритм Нудлмана-Вунша, а для локального – Сміта-Вотермана.

Щодо надзвичайно подібних послідовностей, то між локальним і глобальним вирівнюванням немає різниці.

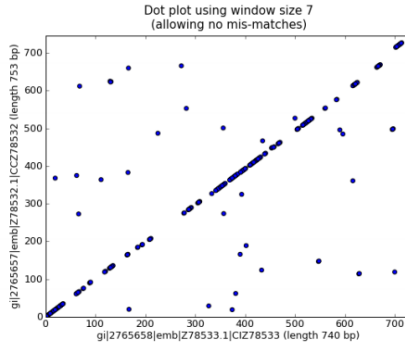
3. Вирівнювання, яке перекривається, призначене для порівняння послідовностей, у яких збігаються тільки кінцеві ділянки.



Культури. Клітин. Метод. Первинної.  
| | | | | | | | | | | | | | | |  
Метод. Вторинної. Гетероплоїдної. Культури. Клітин.

Використовується для збирання послідовностей під час проєктів секвенування геномів.

4. Точкове вирівнювання (dot plot) застосовується для загального дослідження послідовності з метою виявлення повторів і вибору фрагмента для множинного вирівнювання.



**Рис. 16. Схема точкового вирівнювання**

Точкове вирівнювання та вирівнювання, яке перекривається передбачають використання тільки парного порівняння послідовностей.

Зазвичай у біоінформатичних дослідженнях здійснюють порівняння нової послідовності з уже відомими для виявлення спорідненості між ними. Для позначення ступеня спорідненості послідовностей застосовують такі поняття, як «гомологи», «ортологи», «паралоги».

У біоінформатиці два білки називають *гомологічними*, якщо вони за походженням мають спільного предка.

Якщо два білки мають спільного предка та виконують однакові функції у двох різних видів їх називають «*ортологи*». До утворення ортологів призвів акт видоутворення: коли ген існував у виду, який дивергував з утворенням двох видів, то копії цього гена у дочірніх видах називаються ортологами.

«Паралогічні» білки споріднені між собою внаслідок дуплікації їхніх генів усередині генома. Тобто, якщо в межах одного організму в результаті хромосомної мутації відбулася дуплікація гена, то його копії називають *паралогами*.

Отже, ортологи виконують однакові функції в різних організмах, натомість паралоги – різні, але споріднені в одному організмі. Критерієм гомологічності в білках є рівень консервативних послідовностей між цими білками. Для встановлення таких консервативних сайтів виконують множинне вирівнювання.



---

---

## 2. Основні підходи парного та множинного вирівнювання послідовностей

Вирівнювання послідовностей нуклеотидів або амінокислотних залишків зазвичай подають у вигляді рядків у матриці. Між залишками вставляються пропуски так, що залишки з ідентичними або подібними особливостями вирівнюються в послідовних колонках.

Для здійснення вирівнювань використовують алгоритм, який дає змогу стверджувати, що проведене вирівнювання найкраще репрезентує ступінь спорідненості послідовностей. Наприклад є дві послідовності:

```
1: ATGTCGTCAAGGTAATCCA
2: ATGCGTCGGTAATGCT
```

Наше завдання виконати вирівнювання цих послідовностей так, щоб вони були найбільш інформативними. Можливі три різні варіанти вирівнювання цих послідовностей:

### 1. Беззмислове вирівнювання

```
ATGTCGTCAAGGTAATCCA-----
-----ATGCGTCGGTAATGCT---
```

### 2. Вирівнювання без проміжків

```
ATGTCGTCAAGGTAATCCA
ATGCGTCGGTAATGCT---
```

### 3. Вирівнювання з проміжками

```
ATGTCGTCAAGGTAATCCA
ATG-CGTC--GGTAATGCT
```

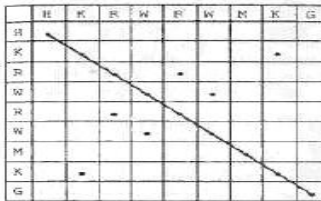
Рамками виділені заміни, які мають значення при визначенні ваги вирівнювання. Як видно, останнє вирівнювання найкраще, оскільки в ньому отримана максимальна кількість збігів для нуклеотидів у двох послідовностях і використана мінімальна кількість вставок.

Закономірність таких записів так само залежить від типу вирівнювання, який використовується.

**Парне вирівнювання послідовностей**

Є три види методів для реалізації парного вирівнювання – точково-матричний, динамічне програмування і так звані «словесні методи».

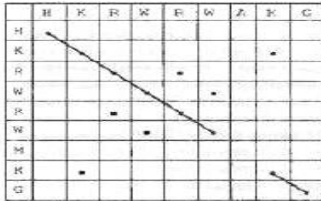
**Точково-матричний підхід** використовують для вирівнювання дуже схожих послідовностей. Для побудови графіка одну з послідовностей записують уздовж верхнього рядка зліва направо, а іншу – вздовж крайньої лівої колонки зверху вниз. Якщо амінокислотний залишок послідовності, розміщений у верхньому рядку, відповідає тому, який міститься в колонці, то в місці перетину ставлять крапку. У результаті виходить двовимірна матриця. Точкові графіки близькоспоріднених послідовностей мають вигляд лінії вздовж діагоналі матриці.



**Послідовності до і після вирівнювання**

```

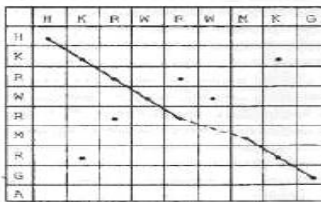
H K R W E W M K G
| | | | |
H K R W M K G
  
```



**Послідовності до і після вирівнювання**

```

H K R W E W L K G
| | | | | * |
H K R W M K G
  
```



**Послідовності до вирівнювання**

```

H K R W M K G
| | | | |
H K R W M K G
  
```

**Послідовності після вирівнювання**

```

H K R W M K G
| | | | |
H K R W M K G
  
```

**Рис. 17. Графіки точково-матричного вирівнювання послідовностей**

---

---

Із точково-матричного графіка дуже просто візуально визначити деякі характерні ознаки послідовності, зокрема інсерції, делеції, повтори чи інвертовані повтори. Наприклад, порушення цілісності діагональної лінії вказує на те, що у вихідній предковій послідовності відбулися інсерції та делеції.

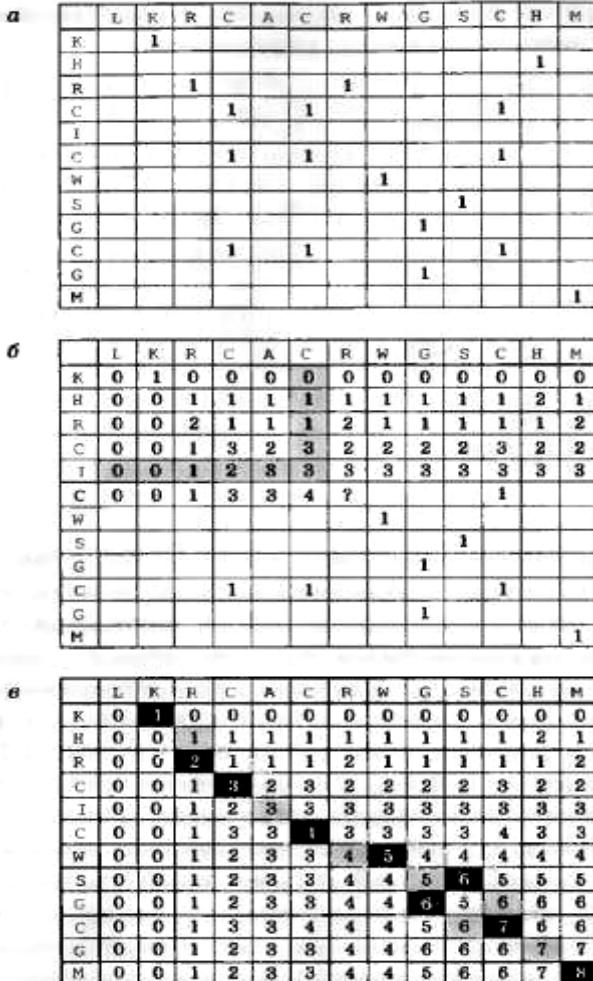
**Метод динамічного програмування** допомагає виконати як локальне, так і глобальне вирівнювання. Цей підхід має три етапи: ініціалізації, заповнення матриці (скоринг) та вирівнювання. На першій стадії створюється матриця з кількістю колонок і рядків, що на 1 більше, ніж кількість мономерів у вирівнюваних послідовностях. Першу колонку і перший рядок матриці заповнюють нулями. Далі заповнення розпочинають із лівого верхнього кутка матриці.

На першому етапі вирівнювані послідовності записуються як перший рядок і перший стовпчик таблиці-матриці, і клітинкам із однаковими амінокислотами призначається значення 1, а клітинкам із різними амінокислотами – 0 (рис. 18а). Після цього визначається максимально можлива сума чисел у всіх клітинках при всіх можливих кроках через матрицю зверху-вниз зліва-направо.

Перший (прямий) прохід матриці. Починаючи із верхньої лівої клітинки, до значення кожної клітинки додається максимальне число із рядка-стовпчика зверху-зліва, як-от, для поміченої знаком питання клітинки до її власного значення (0) додається максимальне число із попередніх рядків-стовпчика, виділених сірим кольором (3), і значення цієї клітинки стає рівним 3 (рис. 18б)

Після заповнення у такий спосіб усіх клітинок матриці, визначений максимально можливий рахунок (у даному разі – 8) (рис. 18в). Рухаючись у протилежному напрямку (від клітинки зі значенням 8), визначається оптимальний шлях проходження матриці. У цьому випадку два шляхи – два варіанти переходу від значення 5 до значення 7 – однаково оптимальні.

У кожному із цих двох альтернативних варіантів вирівнювання у послідовностях збігаються вісім амінокислот, що досягається п'ятьма прогалинами кожна завдовжки в одну амінокислоту.



*Рис. 18. Принцип алгоритму Нідлмана-Вунша для вирівнювання послідовностей*

Розглянутий приклад використання алгоритму Нідлмана-Вунша спрощений: не введені штрафи за прогалини (точніше, вони дорівнюють 0), для будь-яких розбіжностей амінокислот використана однакова, нульова, оцінка.

---

---

Однак введення штрафів та/або різних, більш складних оцінок не змінює зміст алгоритму.

Алгоритм Нідлмена-Вунша спрямований на вирівнювання двох послідовностей по всій довжині. Такі підходи називають спрямованими на глобальне вирівнювання. Принцип таких підходів – максимізація кількості однакових нуклеотидів (амінокислот) та двох послідовностей.

Найуживаніші методи вирівнювання – це так звані *словесні методи*. Вони визначають серію коротких амінокислотних послідовностей («слів»), які не перекриваються в структурі, котра порівнюватиметься із базою даних послідовностей. Ці методи найчастіше використовують у пошукових інструментах баз даних FASTA і BLAST.

В цьому методі використовують алгоритм Сміта-Вотермена. Алгоритм передбачає, що оптимальний шлях через матрицю може починатися та закінчуватися у будь-якій клітинці, а не обов'язково тільки в клітинках першого чи останнього стовпчика. Алгоритм Сміта-Вотермена спрямований не на глобальне, а на локальне вирівнювання.

Продемонструвати принципову різницю в алгоритмах Нідлмена-Вунша та Сміта-Вотермена можна так. Часто необхідно вирівняти не повністю перекриті послідовності. Наприклад, одна із двох вирівнюваних послідовностей має кінець гена 1 та початок гена 2, а друга – повний ген 2 (його початок тікінєць). У цьому разі використання алгоритму Нідлмена-Вунша може призвести до неправильного вирівнювання, що не має біологічного сенсу.

```
кінець_гена_1_початок_гена_2
***** | | | | | * | ***** | | | | | |
початок_гена_2_кінєць_гена_2
```

### *Неправильне вирівнювання*

Тут вирівняні не гомологічні ділянки послідовностей. Використання алгоритму Сміта-Вотермена призведе до виявлення гомологічних ділянок аналізованих послідовностей.

---

```

кінець_гена_1_початок_гена_2.....
      |||||
.....початок_гена_2_кінець_гена_2

```

У практичній роботі не завжди відомо, чи вирівнювані послідовності повністю або частково перекриті.

Розглянемо дві послідовності.

```

1  GAAACACCCTCACAATTGTA
2  ATGGCAGGCTTTACACAT

```

Використання алгоритму Нідлмена-Вунша зумовлює їх вирівнювання у такий спосіб.

```

1  GAAACACCCTC.ACAATTGTA
   ****||**||* |||**|
2  ATGGCAGGCTTTACACAT...

```

Використання алгоритму Сміта-Вотермена для тих самих послідовностей може призвести до такого результату.

```

1  .....TTGACACACTCCCAATTGTA
   ||*|||| |
2  ACCCCAGGCTTTACACA.T.....

```

Число збіжних нуклеотидів в обох випадках однакове – 8. Однак використання алгоритму Нідлмена-Вунша веде до 9 неоднакових нуклеотидів, а використання спрямованого на мінімізацію числа неоднакових нуклеотидів алгоритму Сміта-Вотермена – тільки до одного. При цьому використання алгоритму Сміта-Вотермена зумовлює більше числа прогалин – однак більшість із цих прогалин (усі, окрім однієї) розміщені у неперекритих (можливо негомологічних) ділянках послідовностей.

Отже, використання алгоритму Сміта-Вотермена привело до ідентифікації в обох послідовностях достатньо довгих, подібних між собою, можливо гомологічних одна до одної ділянок.

## Множинне вирівнювання послідовностей

Множинне вирівнювання послідовностей застосовують для одночасного порівняння більш ніж двох послідовностей. Цей вид вирівнювання значно інформативніший, ніж парний, оскільки консервативність низки молекул значно важливіша, ніж повторюваність амінокислотних залишків у двох протеїнів. Консервативні мотиви можна використовувати для визначення каталітично активних сайтів ензимів та еволюційної спорідненості протеїнів.

CLUSTAL 2.1 multiple sequence alignment

```

ssp_330M_2508      -----ATG-----ACTAGT-----TTGTTGTGTTTAA-CGATA----- 28
abc_305D5012_E2791 -----ATG-----ACTAGT-----TTGTTGTGTTTAA-CGATA----- 28
sec_8C2433         -----ATGGGACAGAGAGTGCATATTCAGTCACTGCCTTCCA-CGACA----- 43
etm_5TM2425       -----ATGGGACAGAGACTGCATATTCAGTCACTGCCTTCCA-CGACA----- 42
lta_3L4160        -----ATG-----ATG-----GA-GGAGG----- 10
hza_B366         -----ATG-----ATG-----GA-GGAGG----- 10
lpe_DP1321        -----ATG-----ATGAA-----GCTGGCGCCCCCATCGCA----- 29
xoo_X002023       TTGCGCTCTTCCCAACCCGGCCCGCATAGCGATCGAACCGACAGCCACCTCGT 60
hzu_B3U28020     -----ATC-----TCTATCGATA----- 19
lth_DT_4438      -----ATG-----TATGCAATRAAG----- 16
                    **                               *

ssp_330M_2508      -----AGAG-TAGGGT-GCT----SCAGGCGGATATCGTCCGGT-CCAGTC 68
abc_305D5012_E2791 -----AGAG-CAGGGT-GCT----SCAGGCGGATATCGTCCGGT-CCAGTC 68
sec_8C2433         -----AGCA-TGGGGT-ATG----ACAGACCGGATATCGTCCGGT-CCAGTC 82
etm_5TM2425       -----AGCA-TGGGGT-ATG----ACAGACCGGATATCGTCCGGT-CCAGTC 82
lta_3L4160        -----AGTG-CCGGGT-GCT-----CTCCAT-TCAGAG 33
hza_B366         -----AGTG-CCGGGT-GCT-----CTCCAT-TCAGAG 35
lpe_DP1321        -----GGCCCTTCCCGGGCT-----GCCCATCGAAGCTGTCTCGAT-CCAGTC 71
xoo_X002023       CATGGTCCCGCCGTCGATCCGGATGGCCCTTCGCGGATCGAGTGTATTCGGT-GCAAT 111
hzu_B3U28020     -----AAGT-ACT-----CAGCATTCGGGGT 22
lth_DT_4438      -----TRAGG-AGGATAGCT-----GCCCT-TCA-TG 40
                    * * *

```

**Рис. 19. Множинне вирівнювання нуклеотидних послідовностей**

До методів множинного вирівнювання належать динамічне програмування, прогресивні методи, ітеративні методи та профільний аналіз для знаходження мотивів.

**Динамічне програмування** потребує конструювання  $n$ -вимірного еквівалента матриці, описаній вище для двох послідовностей, де  $n$  – число послідовностей. Для реалізації цього методу потрібні значні комп'ютерні затрати. З метою зменшення комп'ютерних потреб створено пакет програм MSA, який ґрунтується на «сумі пар» цільової функції.

**Прогресивні методи** – це так звані ієрархічні методи, які спочатку порівнюють найподібніші послідовності і поступово додають менш споріднені. Спочатку подібність визначають за допомогою парного вирівнювання, наприклад за використання

---

---

методу FASTA. Результати прогресивного вирівнювання залежать від вибору найбільш споріднених послідовностей на початковому етапі парного вирівнювання. Прогресивні методи вирівнювання використовують для конструювання філогенетичних дерев та передбачення структури протейнів. До програми, яка ґрунтується на прогресивних методах вирівнювання належать Clustal (<http://www.ebi.ac.uk/Tools/clustalw>) і T-Coffee (<http://www.ebi.ac.uk/Tools/t'coffee/index.html>). На зазначених ресурсах можна виконати вирівнювання за допомогою цих підходів.

*Ітеративними методами* спочатку здійснюють глобальне вирівнювання груп послідовностей, а потім – низку повторних вирівнювань для підмножин послідовностей. У процесі повторного вирівнювання ці алгоритми здатні виправляти помилки.

Ці методи можна застосовувати навіть до послідовностей із дуже низьким ступенем подібності.

*Профільний аналіз* виконується в процесі глобального множинного вирівнювання послідовностей, в результаті якого у досліджуваному ряді послідовностей знаходять короткі консервативні мотиви амінокислотних залишків. Такі консервативні регіони ізолюють і використовують для конструювання сітки профільних матриць.

У профільній матриці значення частоти для кожної амінокислоти отримують із її розподілу безпосередньо в консервативних регіонах, а не із загального емпіричного розподілу.

Для множинного вирівнювання послідовностей також застосовують алгоритм загальної оптимізації – приховану модель Маркова. На ньому основана програма FSA (Fast Statistical Alignment). Ця програма спочатку здійснює парне порівнювання послідовностей, а потім використовує технологію відбору послідовностей для реалізації множинного вирівнювання. Цей підхід вирівнювання значно точніший порівняно з вищеописаними завдяки застосуванню алгоритму «крутого підйому». Програму FSA разом з інструментом візуалізації використовують через веб-інтерфейс на сайті: <http://orangutan.math.berkeley.edu/fsa/>, а ресурс кодів доступний через інтернет-сторінку: <http://fsa.sourceforge.net/>.



---

### 3. Програмне забезпечення вирівнювання послідовностей.

#### Міра подібності послідовностей

Зазвичай, після здобуття результатів секвенування (або після завантаження даних із баз Genbank, UniProt) потрібно виконати процедуру вирівнювання нуклеотидних (або амінокислотних) послідовностей. Процедура вирівнювання забезпечує знанням про те, які ділянки ДНК або білка гомологічні, і саме на основі відмінностей у нуклеотидних або амінокислотних замінах гомологів зможемо потім реконструювати філогенетичне дерево.

Програм, які допомагають здійснити вирівнювання секвенованих послідовностей, безліч. Найпоширеніші такі (більшість із них безкоштовна) як:

- ClustalW
- MUSCLE
- T-Coffee
- BLAST
- Mafft
- KAlign
- MEGA

На сьогодні множинне вирівнювання послідовностей реалізується кількома програмами, доступними в режимі онлайн.

Одна з найуживаніших програм прогресивного множинного вирівнювання *ClustalW* – це третє покоління програм цієї серії, які з'явилися в 1994 році. Дана версія значно простіша в роботі завдяки вдосконаленому алгоритму, основаному на створенні множинного вирівнювання в результаті серій попарних вирівнювань, відповідно до відгалужень спрямованого дерева, побудованого методом UPGMA.

Окрім цього, виникла можливість вибирати матриці порівняння амінокислот і нуклеотидів, а також встановлювати штрафи за внесення пропусків. Результати вирівнювання подаються у вигляді формату FASTA, що забезпечує високу сумісність програм цього покоління з іншими пакетами програм.

Останні програми серії Clustal допомагають створювати найбільш коректні множинні вирівнювання біологічних послідовностей. Ця програма дає змогу вирівнювати як нуклеотидні, так і амінокислотні послідовності.

---

Програма доступна на багатьох серверах (<http://www.ebi.ac.uk/services>) у двох варіантах – інтерактивному та поштовому.

Інтерактивний варіант передбачає швидке отримання користувачем результатів вирівнювання (доцільно застосовувати при невеликій (до 100) кількості послідовностей), а поштовий – електронною поштою (при великій кількості послідовностей).

Основне призначення програми ClustalW – виконання множинного вирівнювання, обчислення еволюційних дистанцій між послідовностями, визначення характеру і типу амінокислотних замінів.

Отримані вирівнювання можна відобразити в чорно-білій або кольоровій гамі, залежно від властивостей амінокислот (у разі вирівнювання амінокислот). Консервативність і напівконсервативність амінокислотних замінів визначається відповідно до таблиці 2.

**Таблиця 2**

**Один із можливих способів забарвлення амінокислотних залишків при візуалізації множинного вирівнювання амінокислотних послідовностей**

<b>Колір</b>	<b>Тип залишку</b>	<b>Амінокислоти</b>
<i>жовтий</i>	неполярні залишки	Gly, Ala, Ser, Thr
<i>зелений</i>	гідрофобні	Cys, Val, He, Leu, Pro, Phe, Tyr, Net, Trp
<i>фіолетовий</i>	полярні	Asn, Gin, His
<i>червоний</i>	негативно заряджені	Asp, Glu
<i>синій</i>	позитивно заряджені	Lys, Arg

Якщо амінокислоти розміщуються в одній групі, то заміни вважають консервативними.

На сьогодні найоптимальнішою і найсучаснішою з доступних онлайн-ресурсів є програма MUSCLE (<http://www.ebi.ac.uk/Tools/msa/muscle>). Дане скорочення утворилося від повної назви програми – Multiple Sequence Comparison by Log-Expectation. MUSCLE забезпечує велику точність вирівнювання і вищу

---

швидкість виконання роботи, ніж ClustalW і T-Coffee. Для множинного вирівнювання часто використовують програму MatLab із застосуванням команди «multialign».

Виконати множинне вирівнювання послідовностей можна безпосередньо у базі даних. Зокрема, ресурс поданий в базі UniProt. Однак у ньому є обмеження. Оскільки це база даних амінокислотних послідовностей, то й алгоритми, які реалізують множинне вирівнювання, основані лише на матрицях заміні, а отже, виконати аналіз нуклеотидних послідовностей не можна.

Велика частина програмного забезпечення працює безпосередньо з форматом FASTA.

У біоінформатиці, формат FASTA – це текстовий формат файлу для збереження нуклеотидних або амінокислотних послідовностей, у якому нуклеотиди або амінокислоти позначаються літерами. Цей формат дає змогу передавати опис цих послідовностей і короткий коментар до них. Назва формату походить від програмного пакета FASTA, але він уже став незалежним стандартом у біоінформатиці. Послідовність у цьому форматі починається з назви, перед якою ставлять символ «>». Перше слово після цього символу зазвичай – ідентифікатор послідовності – такий як номер послідовності в базі даних UniProt або GenBank.

```
>sp|Q3TDX8|NB5R4_MOUSE Cytochrome b5 reductase 4 OS=Mus
musculus GN=Cyb5r4 PE=2 SV=3
MLNVPSQAFAPAGSQQRVSSQGRSKVPLKQGRSLMDWIRLTKSGKDLTGLKGGGLIEVTEE
ELKKHNKKEDCWICIRGFVYNVSPYMEYHPPGGEDELMRAAGADGTDLFNEVHRWVNYESM
LKECLVGRMAVYKPAVPKDCHEGKRVLNGMLPKSQMSDTLPRDVTDTLPREDLSSPSYDWF
QTESVTVIVVYTKQKNISLDSVIVDLQDDSLRAEAVIKDHSYLVHVGLSHEVQENFSVRV
IENVGKIEIVLQKKEVSVWQCLGDHLEKHSFIPKKDTGLYYRRCQLISKEDVTHDTRLF
CLMLPPSTHLQVPVQGHVYLKLSVTGAEIVKPYTPVSDSLLSDFKPEVLSPNKYICFLIK
IYPAGLFTPELDRLQIGDFISVSGPEGDFKVSKLQEVEDLFLAAGTGFTPMVTVLNLYAL
SHMSSLRKVKLMFFNKTEDDIIWRCQLEKALREKRFDFEVLVLSAPSPPEWNGKQGHISRA
LLSEFLQRSSENSRAFLCICGPTPFTDEGIRLLHDLNFSDEIHHGFTA
```

**Рис. 20. Приклад формату FASTA амінокислотної послідовності**

Останні слова в першому рядку можуть передавати будь-яку інформацію про послідовність. Усі слова в першому рядку

---

---

необов'язкові та ймовірні у довільному форматі. Потім із нового рядка вводять саму послідовність. Формат рекомендує обмежувати довжину рядків до 80 символів. Найчастіше рядки послідовності мають довжину 60 символів, інколи, різну – це межа зі «рваним» правим краєм.

У FASTA-форматі використовують однобуквені коди для нуклеотидів і амінокислот, задані Міжнародним об'єднанням біохімії і Міжнародним об'єднанням чистої і прикладної хімії.

#### *Міра подібності послідовностей*

Для більшої впевненості у правильності вирівнювання розроблено показники вимірювання подібності та різниці між послідовностями, до яких належать:

1. Відстань Гемінга – кількість позицій (мономерів) у ланцюгах однакової довжини, за якими вони відрізняються;

2. Відстань Левенштейна (відстань редагування) – мінімальна кількість операцій редагування (вирізання, вставка чи заміна) між двома ланцюгами необов'язково однакової довжини, необхідних для того, щоб зробити один ланцюг ідентичним до іншого.

AGTC CGTA	Відстань Гемінга = 2
AG-TCC CGCTCA	Відстань Левенштейна = 3

Відстань Гемінга і Левенштейна – міра подібності двох послідовностей: подібні послідовності дають низькі відстані, а неподібні – великі.

#### **4. Типи та спосіб побудови філогенетичних дерев**

Філогенетичний аналіз у систематиці описує взаємозв'язки між таксонами і допомагає зрозуміти історію еволюційних відносин між живими організмами. Еволюційну історію, відновлену в результаті філогенетичного аналізу, зображають у вигляді розгалужених, деревоподібних діаграм, які подають передбачуваний родовід спадкових взаємозалежностей між молекулами, організмами або і тим, і тим.

У філогенетиці найзручніший спосіб візуального подання еволюційних зв'язків серед груп організмів здійснюється за допомогою графіків, які називаються філогенетичними деревами.

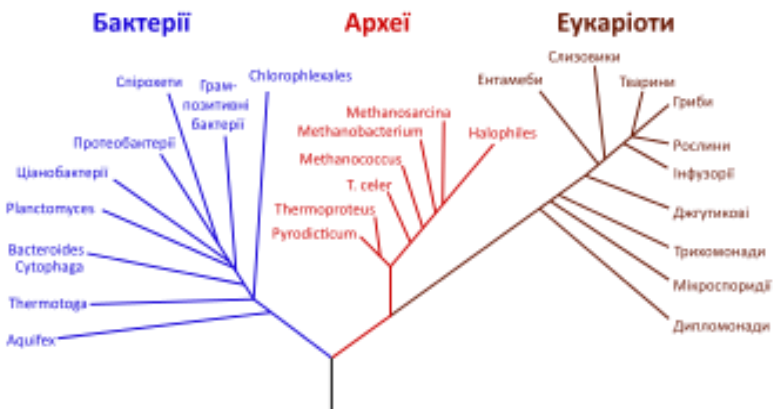
*Філогенетичне дерево* – дерево, яке відображає еволюційні взаємозв'язки між різними видами, іншими таксонами, генами, білками або іншими об'єктами, котрі мають загального предка.

Вершини філогенетичного дерева поділяються на три класи: листя, вузли і (максимум один) корінь. Листя – це кінцеві вершини; кожен лист відображає деякий вид живих організмів (або інший об'єкт, схильний до еволюції, наприклад білковий домен). Кожен вузол – це еволюційна подія: поділ предкового виду на два або більше, які надалі еволюціонували незалежно. Корінь представляє загального предка всіх даних об'єктів. Ребра дерева філогенезу прийнято називати «гілками».

Ідея «дерева» з'явилася в ранніх поглядах на життя як на процес розвитку від простих форм до складних. Сучасні еволюційні біологи використовують дерева для ілюстрації еволюції, оскільки воно наочно показує розвиток і походження видів.

Розрізняють два види філогенетичних дерев: укорінене та неукорінене.

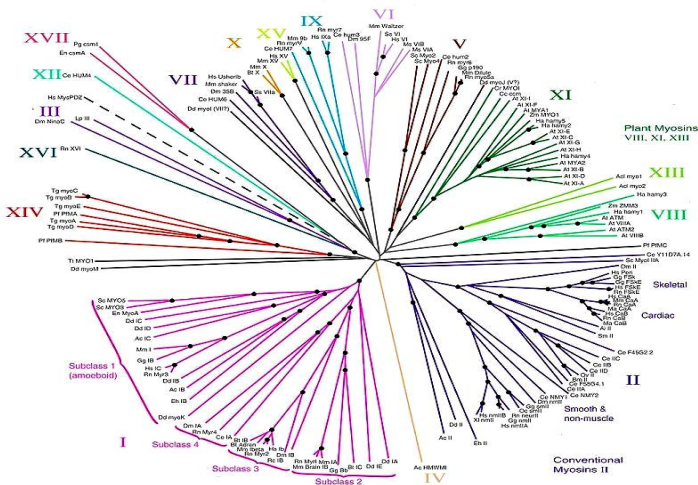
Укорінене дерево – дерево, із виділеним коренем.



*Рис. 21. Укорінене філогенетичне дерево*

Укорінене дерево можна вважати орієнтованим графіком, оскільки воно має природну орієнтацію – від кореня до листя. Кожен вузол укоріненого дерева відповідає останньому загальному предкові листя дерева, яке лежить нижче. На малюнку представлене вкорінене філогенетичне дерево, забарвлене відповідно до доменів живих організмів у «системі трьох доменів».

Невкорінене дерево не містить кореня і відображає зв'язок листя без передбачуваного положення загального предка. Необхідність розглядати невикорінені дерева виникає тому, що часто зв'язки між вузлами відновити легше, ніж напрямок еволюції. На рис. 22 показано невикорінене філогенетичне дерево.



**Рис. 22. Невкорінене філогенетичне дерево**

Найдостовірніший метод для перетворення невикоріненого дерева на вкорінене (для цього треба або вибрати коренем один із вузлів, або розбити одну з гілок на дві, які виходять із кореня) – використання достовірної «зовнішньої групи» видів – достатньо близьких до набору видів, який цікавить нас, але водночас явно є окремою групою.

Як укорінене, так і невикорінене філогенетичні дерева можуть бути біфуркаційними або мультифуркаційними, а також маркованими чи немаркованими. У біфуркаційному дереві до

---

---

кожного вузла підходять рівно три гілки (у разі вкоріненого дерева – одна вхідна гілка і дві вихідні). Тобто біфуркаційне дерево припускає, що всі еволюційні події полягали в походженні рівно двох нащадків від предкового об'єкта.

До вузла мультифуркаційного дерева можуть підходити чотири та більше гілок. Марковане дерево містить назви листків, натомість немарковане просто відображає топологію.

Завжди є більше можливих мультифуркаційних дерев, ніж біфуркаційних, більше маркованих, ніж немаркованих і більше вкорінених, ніж невикорінених. Остання відмінність має найбільше біологічне значення – є багато місць на невикоріненому дереві, куди можна помістити корінь.

За наявністю довжини гілок у філогенетичних деревах їх поділяють на кладограми, філограми, хронограми.

*Кладограма* – філогенетичне дерево, яке не містить інформації про довжини гілок.

*Філограма (або фенограма)* – філогенетичне дерево, яке містить інформацію про довжини гілок – ці довжини демонструють зміну якоїсь характеристики.

*Хронограма* – філограма, довжини гілок у якій вказують на еволюційний час.

*Дендрोगрама* – загальний термін на позначення схематичного відображення філогенетичного дерева.

Філогенетичні дерева, які складаються з нетривіального числа вхідних послідовностей, побудовані з використанням обчислювальних філогенетичних методів.

Для виконання подібних філогенетичних реконструкцій зручно використовувати онлайн-програми. Найзручніший для студентів пакет програм «Robust Phylogenetic Analysis For The Non-Specialist» на сайті <http://www.phylogeny.fr/>.

Phylogeny.fr – це безкоштовний, простий у використанні веб-сервер, присвячений реконструкції і аналізу філогенетичних зв'язків між біологічними послідовностями. В основі роботи сервера Phylogeny.fr лежить об'єднання різних біоінформаційних програм для реконструкції надійного філогенетичного дерева з набору послідовностей. Багато методів вирівнювання послідовності, як-от ClustalW, BLAST, також створюють дерева за допомогою простих алгоритмів у побудові дерева.

# ПРАКТИКУМ

## Завдання 1

Виконайте вирівнювання амінокислотної послідовності супероксиддисмутازی людини проти амінокислотної послідовності відповідного білка щура, дрозофіли, кишкової палички. Побудуйте філогенетичне дерево. Вкажіть найспорідненіші види.

## Рекомендація

1. Здійснюємо вхід на веб-ресурс UniProt (<http://www.uniprot.org>). У полі пошуку вводимо запит для супероксиддисмутازی людини (human superoxide dismutase) і натискаємо «Search».

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

- UniProtKB** (555,388): Manually annotated and reviewed. Records with information extracted from literature and curator-evaluated computational analysis. TrEMBL (102,248,261): Automatically annotated.
- UniRef**: The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.
- UniParc**: UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.
- Proteomes**: A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

**Supporting data**

- Literature citations
- Taxonomy
- Subcellular locations

**News**

- Forthcoming changes
- Planned changes for UniProt
- UniProt release 2017\_12: Swiss-Prot in the sky with psilocybin: the biosynthesis pathway of a psychedelic drug unveiled
- UniProt release 2017\_11: Site determination in insects: 50 ways to achieve sex-specific splicing

2. Із результатів пошуку вибираємо той, який відповідає супероксиддисмутазі людини (наприклад з ідентифікатором P00441).

**UniProtKB results**

Filter by' BLAST Align Download Add to basket Columns 1 to 25 of 822

Quote terms: "superoxide dismutase"

Entry	Entry name	Protein name	Gene names	Organism
P00441	SODC_HUMAN	Superoxide dismutase [Cu-Zn]	SOD1	Homo sapiens (Human)
P09487	A4_HUMAN	Amyloid-beta A4 protein	APP, A4, A41	Homo sapiens (Human)
P08228	SODC_MOUSE	Superoxide dismutase [Cu-Zn]	Sod1	Mus musculus (Mouse)
P04179	SODM_HUMAN	Superoxide dismutase [Mn], mitochon...	SOD2	Homo sapiens (Human)
P06442	SODC_BOVIN	Superoxide dismutase [Cu-Zn]	SOD1	Bos taurus (Bovine)
Q99497	PARK7_HUMAN	Protein/nucleic acid deglycase B3-1	PARK7	Homo sapiens (Human)

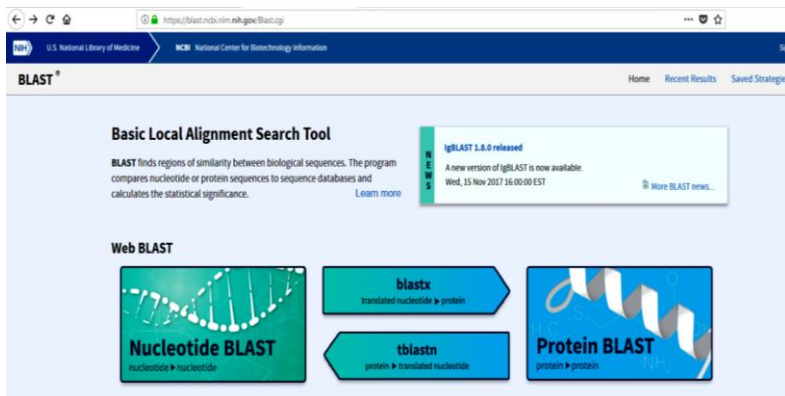


3. Заходимо на цей результат пошуку і знаходимо амінокислотну послідовність. Натискаємо «FASTA».

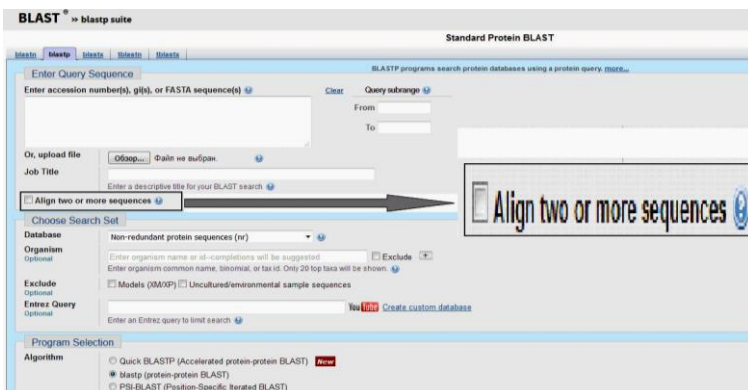
4. Отримуємо амінокислотну послідовність у форматі FASTA, яку зберігаємо у задалегідь створеному файлі.

5. Кроки алгоритму 1 – 4 виконуємо для супероксиддисмутази щура (rat superoxide dismutase), дрозофіли (*Drosophila melanogaster* superoxide dismutase), кишкової палички (*Escherichia coli* superoxide dismutase).

6. Здійснюємо вирівнювання послідовностей із використанням програми BLAST. Для цього заходимо на веб-ресурс <https://blast.ncbi.nlm.nih.gov> і отримуємо сторінку з вікном програми. Вибираємо опцію «Protein BLAST», оскільки ми виконуємо вирівнювання амінокислотних послідовностей.



7. Під час множинного вирівнювання амінокислотних послідовностей вибираємо функцію «Align two or more sequences».



8. Відкривається два поля. В перше поміщаємо амінокислотну послідовність супероксиддисмутази людини, у друге – послідовності супероксиддисмутази щура, дрозофіли, кишкової палички. Натискаємо «BLAST».

NIH U.S. National Library of Medicine NCB I National Center for Biotechnology Information

**BLAST** » blastp suite

Align Seq

blastn blastp **blasts** tblastn tblastx

Enter Query Sequence BLASTP program search pr

Enter accession number(s), gi(s), or FASTA sequence(s) Clear From subrange

MATKAVCVLKGSDGPFVQGIINFEQKESNGPVKVGSIKGLTEGLGHFVHFEGDNTAGCTS  
 AGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV  
 HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ

From:   
 To:

Or, upload file  Файл не выбран.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Subject subrange

MATKAVCVLKGSDGPFVQGIINFEQKESNGPVKVGSIKGLTEGLGHFVHFEGDNTAGCTS  
 AGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV  
 HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ

From:   
 To:

Or, upload file  Файл не выбран.

Program Selection

Algorithm  blastp (protein-protein BLAST)  
 Choose a BLAST algorithm

Search protein sequence using Blastp (protein-protein BLAST)  
 show results in a new window

## 9. Отримуємо результат множинного вирівнювання.

Alignments

Download Graphics

sp|P07632|SODC\_RAT Superoxide dismutase [Cu-Zn] OS=Rattus norvegicus GN=Sod1 PE=1 SV=2  
 Sequence ID: Query\_88645 Length: 154 Number of Matches: 1

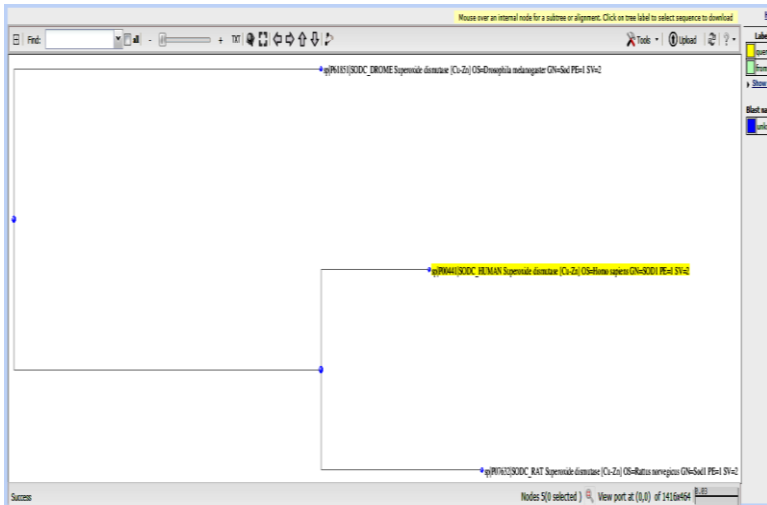
Range 1: 1 to 154	Graphics	Score	Expect	Method	Identities	Positives	Gaps
288 bits(660) 2e-94 Compositional matrix adjust. 128/154(83%) 137/154(88%) 0/154(0%)							
Query 1		MATKAVCVLKGSDGPFVQGIINFEQKESNGPVKVGSIKGLTEGLGHFVHFEGDNTAGCTS			60		
Sbjct 1		MA KAVCVLKGSDGPFVQGIINFEQKESNGPVKVGSIKGLTEGLGHFVHFEGDNTAGCTS			60		
Query 61		AGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV			120		
Sbjct 61		AGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV			120		
Query 121		HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ			154		
Sbjct 121		HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ			154		

Download Graphics

sp|P61851|SODC\_DROME Superoxide dismutase [Cu-Zn] OS=Drosophila melanogaster GN=Sod PE=1 SV=2  
 Sequence ID: Query\_88646 Length: 153 Number of Matches: 1

Range 1: 1 to 152	Graphics	Score	Expect	Method	Identities	Positives	Gaps
186 bits(471) 1e-65 Compositional matrix adjust. 94/154(61%) 113/154(73%) 2/154(1%)							
Query 1		MATKAVCVLKGSDGPFVQGIINFEQKESNGPVKVGSIKGLTEGLGHFVHFEGDNTAGCTS			60		
Sbjct 1		M VTKAVCVINGD--AKGTVFFEQSSGTPVKVSGEVGLKAGLGHFVHFEGDNTAGCTS			58		
Query 61		AGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV			120		
Sbjct 59		SGPFPNPLSKHGSGPDEERFVQDLGNVADKDGVAIVSIEDRVISLSGEHSIIIGRTIVV			118		
Query 121		HEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ			154		
Sbjct 119		HADDDLGGHLESTKTNAGSRLACGVIGIAQ			152		

10. Для побудови філогенетичного дерева вибираємо опцію «Reset Tree». Отримуємо філогенетичне дерево. З результатів видно, що найспорідненішими є людина та щур. Кишкова паличка не має спільного предка з цими організмами.



## Завдання 2

На основі амінокислотної послідовності аспартатамінотрансферази людини за допомогою інтернет-ресурсів визначте молекулярну масу та ізоелектричну точку даного ферменту.

### Рекомендація

1. Знаходимо амінокислотну послідовність у FASTA-форматі для аспартатамінотрансферази людини (human aspartate aminotransferase) через веб-ресурс UniProt (див. завдання 1).

```

www.uniprot.org/uniprot/P00505.fasta

>p|P00505|AATM_HUMAN Aspartate aminotransferase, mitochondrial OS=Homo sapiens
MALLHSRGLRPGIAAÄFHFGLAÄASARASSWHTVEMGPPDFILGVTEAFKRDTNSKIM
NLGVGAYRDDNGKPYVLPVSRKAEAQIAAKNLDEYLPIGGLAEFCKASAEALGENSEV
LKSGRFVTVQTIISGTGALRIGASFLQRFFKFSRDVFLPKPTWGNHTPIFRDAGMQLGGYR
YYDPKTCGFDFTGAVEDISKIPEQSVLLHACAHNFTGVDERPEQWKEIATVVKRNLF
FEDMAYQSFASGGDKDAAVRFIEQGINVCLCQSYAKNMGLISERVGAFTMVKRDADE
ARKVESGLTILIRPMYSNPTMGARITAAAILNTPDLRQWLGEVYVMDRIIGMRTQIVS
NLKKEGSTHNQHITDQIGMFCFTGLKPEQVERLIKEFSIYMTKDRISVAGVTTSSNVGY
LAHAIHQVTK

```

2. Здійснюємо вхід на веб-ресурс ExPASy (<https://www.expasy.org>). У меню головного пошукового екрана зліва вибираємо опцію «proteomics».

ExPASy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e. sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.) (see [Categories](#) find resources from many different SIB groups as well as external institutions.

**How to use this portal?**

- Features and updates
- New to ExPASy
- Empowered ExPASy users: what is different?

3. Из запропонованого меню вибираємо ресурс «Compute pI/Mw».

**Resources A..Z**

- Compute pI/Mw**

4. У поле ресурсу вставляємо скопійовану амінокислотну послідовність і натискаємо «Click here to Compute pI/Mw».

**Compute pI/Mw tool**

Compute pI/Mw is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) (k reference).

Documentation is available.

Compute pI/Mw for Swiss-Prot/TrEMBL entries or a user-entered sequence

Please enter one or more UniProtKB/Swiss-Prot protein identifiers (ID) (e.g. *ALBU\_HUMAN*) or UniProt Knowledgebase or a protein sequence in single letter code. The theoretical pI and Mw (molecular weight) will then be computed.

Or upload a file from your computer, containing one Swiss-Prot/TrEMBL ID/AC or one sequence per line.

**Click here to compute pI/Mw**

3. Отримуємо результат. Встановлено, що ізоелектрична точка та молекулярна маса аспаратамінотрансферази людини становлять 47517,65 та 9,14 відповідно.

**Compute pI/Mw**  
**Theoretical pI/Mw (average) for the user-entered sequence:**

10	20	30	40	50	60
MALLHSGRVLI	EGIAAAFHPG	LAAASARAS	SWWTHVEMGF	EDPILGVTEA	FKRDINSKMN
70	80	90	100	110	120
NLGVGAYRDD	NGKPYVLEPV	RKAEAQIAAK	NLDKEYLPIG	GLAEFCRASA	ELALGENSEV
130	140	150	160	170	180
LKSGRFVTVQ	TISGTGALRI	GASFLQRFRR	FSRDVFLPKP	TWGNHTPIFR	DAGMQLQGYR
190	200	210	220	230	240
YYDKFCGGFD	ETGAVEDISK	IPEQSVLLIH	ACAHNPTGYD	EREQWKEEA	TVVKKRNLFA
250	260	270	280	290	300
FFDMAYQGF	SGDGDKDANA	VRHFIEQGIN	VCLCQSYAKH	MGLYGERVGA	FTMVCKDADE
310	320	330	340	350	360
AKRVESQLKI	LIRPHYSNPF	LNGARIAAAI	LNTFDLRKQW	LQEVKVMADR	IIGMRTQLVS
370	380	390	400	410	420
NLKEEGSTNH	WQHITDQIGM	FCFTGLKPEQ	VERLIKEPFI	YMTKDGRIIV	AGVTSSNVGY
430					
LSATHGQKSE					

**Theoretical pI/Mw: 9.14 / 47517.65**

### Завдання 3

Виконайте парне вирівнювання молекули хлорофілу різущки Таля (*Arabidopsis thaliana*) та гороху посівного (*Pisum sativum*). Визначте полярні амінокислоти.

### Рекомендація

1. Дане вирівнювання можна здійснювати в базі даних UniProt (<http://www.uniprot.org>). Для цього у полі пошуку вводимо запит для хлорофілу (chlorophyll) та натискаємо «Search». Отримуємо результати пошуку. Позначаємо ID потрібних організмів відповідними значками та натискаємо «Align».

**UniProtKB results**

Filter by: A: thaliana (397)

Popular organisms: Rice (215), CHRE (141), PEA (66), SYNV3 (45), Other organisms

Search terms: Filter "chlorophyll" as:

Entry	Entry name	Protein names	Gene names	Organism	Length
Q95377	CB21_ARATH	Chlorophyll a-b binding protein 2.2...	LHC2.2 At2g05070, F1013.20	Arabidopsis thaliana (Mouse-ear cress)	265
Q95377	CB21_ARATH	Chlorophyll a-b binding protein 2.3...	LHC2.1 LHC2.3, At2g05100, F15111.2	Arabidopsis thaliana (Mouse-ear cress)	258
Q95377	CB21_ARATH	Chlorophyll a-b binding protein 2.4...	LHC2.4 At3g27690, MGF30.9	Arabidopsis thaliana (Mouse-ear cress)	266
Q95397	LHC3_ARATH	Photosystem I chlorophyll a/b-binding protein	LHC3A At1g61520, T25824.12	Arabidopsis thaliana (Mouse-ear cress)	273
P07371	CB22_PEA	Chlorophyll a-b binding protein ABB...	ABB0	Pisum sativum (Garden pea)	269
Q95397	LHC3_ARATH	Chlorophyll a-b binding protein 2.5...	LHC3.4 At3g27410, PUA3.9	Arabidopsis thaliana (Mouse-ear cress)	267
Q95377	CB1B_ARATH	Chlorophyll a-b binding protein	LHC1.2 AB100, CAB1, LHC1-A, At3g29510, F1N18.5	Arabidopsis thaliana (Mouse-ear cress)	267

## 2. Отримуємо вирівнювання послідовностей.

www.uniprot.org/align/A20180125AAF87E4D2F1D05654627429E83DASCEAC39C

**Alignment**

How to print an alignment in color

Q957J7	CB22_ARATH	1	MAISA-----IQSSFAQ-DTALFSSDILQKVVGLGGGVTRERTYK-----STPQSIWY	50
F07371	CB22_PEA	1	MAASSSSMALSSPFLAGRLKLN-----SQQLGAARFTRRSATTKVRSVSSGSPWY	54
Q957J7	CB22_ARATH	51	GFDRFVYLGFSSEMTFVYLIGEYFGDYQWDTAGLSADPETFAKRRLELVHSRWAMLGAL	110
F07371	CB22_PEA	55	GFDRVYLYGFSGSSPFTYIIGEYFGDYQWDTAGLSADPETFAKRRLELVHSRWAMLGAL	114
Q957J7	CB22_ARATH	111	GCTFPELLSKNGVFGKAVWFKAGSQIFSEGGLDYLGNPLIHAQSILAIWAVQVLMVGF	170
F07371	CB22_PEA	115	GCVFPELLSRNGVVFGEAVWFKAGSQIFSEGGLDYLGNPLVHAQSILAIWAVQVLMVGF	174
Q957J7	CB22_ARATH	171	IEGYIQQGGLFEGLDLPLYGGAFFDLNLADEFAFSLVKKVLENGRGLAMFMSFGFPVQ	230
F07371	CB22_PEA	175	VEGYIAGGGLFGEVVDLYFGGDFPLGLADDFEAFELVKKVLENGRGLAMFMSFGFPVQ	234
Q957J7	CB22_ARATH	231	AIYVGRGPIEMLFDRHADPVAANNMAYATNFVGR	265
F07371	CB22_PEA	235	AIYVGRGPIEMLFDRHADPVAANNMAYATNFVGR	269

You may add additional sequences to this alignment (in FASTA format)

**Amino acid properties**

- Similarity

3. Для визначення полярних амінокислот у лівому меню вибираємо «polar». У результатах ці амінокислоти виділяються фіолетовим кольором.

www.uniprot.org/align/A20180125AAF87E4D2F1D05654627429E83DASCEAC39C

**Alignment**

How to print an alignment in color

Q957J7	CB22_ARATH	1	MAISA-----IQSSFAQ-DTALFSSDILQKVVGLGGGVTRERTYK-----STPQSIWY	50
F07371	CB22_PEA	1	MAASSSSMALSSPFLAGRLKLN-----SQQLGAARFTRRSATTKVRSVSSGSPWY	54
Q957J7	CB22_ARATH	51	GFDRFVYLGFSSEMTFVYLIGEYFGDYQWDTAGLSADPETFAKRRLELVHSRWAMLGAL	110
F07371	CB22_PEA	55	GFDRVYLYGFSGSSPFTYIIGEYFGDYQWDTAGLSADPETFAKRRLELVHSRWAMLGAL	114
Q957J7	CB22_ARATH	111	GCTFPELLSKNGVFGKAVWFKAGSQIFSEGGLDYLGNPLIHAQSILAIWAVQVLMVGF	170
F07371	CB22_PEA	115	GCVFPELLSRNGVVFGEAVWFKAGSQIFSEGGLDYLGNPLVHAQSILAIWAVQVLMVGF	174
Q957J7	CB22_ARATH	171	IEGYIQQGGLFEGLDLPLYGGAFFDLNLADEFAFSLVKKVLENGRGLAMFMSFGFPVQ	230
F07371	CB22_PEA	175	VEGYIAGGGLFGEVVDLYFGGDFPLGLADDFEAFELVKKVLENGRGLAMFMSFGFPVQ	234
Q957J7	CB22_ARATH	231	AIYVGRGPIEMLFDRHADPVAANNMAYATNFVGR	265
F07371	CB22_PEA	235	AIYVGRGPIEMLFDRHADPVAANNMAYATNFVGR	269

You may add additional sequences to this alignment (in FASTA format)

**Amino acid properties**

- Similarity
- Hydrophobic
- Negative
- Positive
- Aliphatic
- Polar

Add sequence and align

**Result information**

### Завдання 4.

Ідентифікуйте білок та його видову належність за послідовністю.

MLNVPSQAFPAPGSQQRVSSQGRSKVPLKQGRSLMDWIRLTKSGKDLTGLKGGLEVTEE  
ELKKNKEDCWICIRGFVYVNSPYMEYHPGGEDELMRAAGADGDTLDFNEVHRVWVNESM  
LKECLVGRMAVKPAVPKDCHEGKRVLNGMLPKSQMSDTLPRDVTDLTPREDLSSPSYDW  
FTQESSVTIVVYTKQKNISLDSVIVDLQDDSLRAEAVIKDHSYLVHVGLSVEQNFVVRV  
IENVGKIEIVLQKKEVSWSQCLGDHLEKHDLSFIPKDDTGLYYRRCQLISKEDVTHDTRLF  
CLMLPSTHLQVPPVGGHVYLLKLSVTGAIEVFPYTPVSDLLSDFKFPVLSPNKYCIFLK  
IYPAGLFTPELDRLLQVGFHFVSVSGPEGDFKVSKLQEVEDLFLLAAGTGFQTFMVTYCNLYAL  
SHMSSLRKVKLMFFNKTEDDIIWRQCLEKALREKRFDFVEFVLSAPSPWNGKQGHISRA  
LLEFLQRSSENSRAFLCICGPTPFDEGIRLLHDLNFSDDIEHGFTA







## Завдання 5

За допомогою програми ClustalW виконайте множинне вирівнювання нуклеотидних послідовностей каталази людини (human), гелікобактера пилорі (*Helicobacter pylori*), грибка ризостонії (*Rhizoctonia solani*).

### Рекомендація

1. Заходимо на веб-ресурс GenBank – <http://www.ncbi.nlm.nih.gov/Genbank>. Знаходимо нуклеотидні послідовності у FASTA-форматі для потрібних організмів.

NCBI Resources How To

Nucleotide Nucleotide catalase human Search

Create alert Advanced

Species Summary - 20 per page Sort by Default order Send to Filters: Manage Filters

Animals (444)  
Plants (183)  
Fungi (1,980)  
Protists (75)  
Bacteria (55,587)  
Archaea (321)  
Viruses (21)  
Customize ...

Molecule types  
genomic DNA/RNA (58,508)  
mRNA (141)  
rRNA (1)  
Customize ...

Source databases  
INSDC (GenBank) (51,954)  
RefSeq (6,703)  
Customize ...

Genetic

See CAT2 (**CATALASE**) **catalase\_2** in the Gene database  
**catalase** reference sequences [Transcript \(3\)](#) [Protein \(3\)](#)

Results by taxon  
Top Organisms [TRe](#)  
[Mycobacterium abs](#)  
[Escherichia coli \(30\)](#)  
[Salmonella enterica](#)  
[Staphylococcus aur](#)  
[Pseudomonas aerug](#)  
[All other taxa \(45885\)](#)  
More ...

Items: 1 to 20 of 58673

of 2934 Next > Last >>

**Human fibroblast catalase gene, partial exon 1, complete exon 2**

1. 1,848 bp linear DNA  
Accession: K02400.1 GI: 940256  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

2. **Human gene for catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)**

Find related data  
Database: Select  
Find items

2. Створюємо файл із нуклеотидними послідовностями, які хочемо проаналізувати, та зберігаємо їх у форматі FASTA.

FASTA ->

### Human fibroblast catalase gene, partial exon 1, complete

GenBank: K02400.1

[GenBank](#) [Graphics](#)

```
>K02400.1 Human fibroblast catalase gene, partial exon 1, complete exon 2
AAAGGGAGCAGGCGCCTTGGCTACTTTGAGGTCACACATGACATTACCAAATACTCCGAGCGAAAGGTA
TTGAGCATATGGAAAGAAGACTCCCATCGCAGTTCGGTTCCTCCTGTTGCTGGAGAATCGGGTTCAGC
TGACACAGTTCGGGCGCCTCGTGGGTTGCGAGTGAATTTTACACAGAAGATGGTAACCTGGGATCTCGTT
GGAAATAACACCCCATTTTCTTCAGGGGATCCCAATATGTTCCCATCTTTATCCACAGCGAAAGA
GAAATCCTCAGACACATCTGAAGGATCCGGACATGTCTGGACTCTGGAGCCATCGCTCGACTCT
GCATCAGTTCCTTTCTGTCAGTGAATCGGGGATTCAGATGGACATCGCCACATGAATGGATGGGA
TCACACACTTTCAGACTGGTTAATGCAAAATGGGGAGCAGTATTATGCAAAATCCATATAAAGACTGACC
AGGGCATCAAAAACCTTCTGTTGAAGATGCGCGAGACTTCCAGGAAGATCCTGACTGGGCATCCG
GGATCTTTTAAAGCCATTGCCACAGGAAAGTACCCCTCTGGACTTTTACATCCAGGTCATGACATT
AATCGGCGAGAAGCTTTCATTTAATCCATTCGATTCACCAAGTGGAGTCACTAAACAACTATATGTTGT
TTCTTTTTTGTAGTCTGCTACCTATTTAGAAAATAATCTAGTCAAGACATTTATATAATGGGGAG
TCATATCAAAAATACAGAGGGTACCCTTCAGAGTGTCTCAAGCTGTAATGAGTGCATACCAGCATCT
ACTTCCAGTCTCGTGTGTGTCATTCATTGAGTATGTGTATGGCTTCATATATGTTATTAACAGGGGA
ACAGATTATGAAAGCTGGATCTGTTCTCGTGGGAAGACTGCAGTATTTACACTACTATTTGTGAAA
GATTTAACTAAGGCATCATTAATCTTATGTTTTATGGATTAATAAATATTTTCATTTGGCTTGA
TTGATTTGAAATCTGGTATTTTGTGGGTAGCTTTGATTTCCCTCAGTGTATTCGCTGGTAAATGTGAA
TATGACATCATTTCAGTTTGGCTCAGCAGGACTCCCTCATCCGAGTGGTAACTGGTCTTAA
CCGGATCCAGTAAATACCTTCTGCTGAGTGTGAACAGATAGCCCTTCGACCCAAGCAACATGCCACCTGCC
ATTGAGGCCATCTCAGCAAAATGCTTCAGGGCCGCCCTTTGCTATCCTCGACACTACCGCCATCGCC
TGGGACCCAAATATCTCATATACCTGTGAAGTGTCCCTACCTGCTCGAGTGGCCAACTACAGCTGA
CGGCCGATGTGCATCGAGGACAAATCAGGGTGGTGTCTCAAATATGACCCCAAGCTTTGGGCGCG
GAACACAGCCTTCTGCCTTGGAGCACAGCATCCAATATTCGGGAAGTGGGGAGATCAACACTGCCA
ATGATGATACCTTACTCAGTTCGGGCACTTATGTGAACGTGCTGAATGAGGACAGAGGAAACTCT
GTGTGGAGACATTCGGGCCACTGTGAAGGATGACAAATTTCTATCCGAGAGAACGGTCAAGACTTC
ACTGAGGTCCACCCTGACTACGGGAGCCATCCAGGCTCTTTCGACAGTACAATCTGGAAGSCTA
AGAATCGGATTCACACCTTTGTGCACTCCGGATCTCACTTGGCGGCAAGGGAGAAAGCAAAATCTGTGAG
CCGGGGCTGCACCTTGTGCATGAAGCT
```

## Helicobacter pylori catalase (katA) gene, complete cds

GenBank: U67458.1

[GenBank](#) [Graphics](#)

```
>U67458.1 Helicobacter pylori catalase (katA) gene, complete cds
AGCATGGATGATGAAATGATGATTTATTTAAACACAGATAAATAAATTTGGTTTTAAATCT
TCTTTATTTTCATTAATTTGTTACGAATAGAAATACTTAAGGGGGTTTTTCATTCTTAAAAAAGGATTT
TTAAGGAATTTGAATCTTGTAGCTTTATATAACAATAATGTAATAATACCACAAGATTAATCGGGTTC
GTGTACAGATTGCGAAATCTAAATCAAATAAAGGAAGAAGATGTTAATAAAGATGTGAACAACCCAC
TGCTTTTGGCGCTCCCGTTTGGGATGATAACAATGTGATTACGGCTGGTCTAGAGGTCCCTGTTTTATA
CAAAGCACTTGGTTTTGGAAAAGTTAGCAGCGTTTGACAGAGAAGGATCCCTGAAAGGGTAGTGATG
CTAAAGGAAGCGGGGCTTATGGCACTTTCACCGTACTAAAGACATCACTAAATACACTAAAGCGAAGAT
TTTTTTAAAGTGGGCAAAAAACCGAATGCTTTTTCAGGTTTTTTACTGTGGCTGGTGAAGAGGCGAGT
CGGGATGCAAGTGAAGACCTTAGAGTTTTGCGATGAAGTATTAACACTGAAGAGGTAACTGGGATTTAG
TAGGAAACAACACGCGCTTTTTCTTATCCGATGATGCGATCAAAATCCCTGATTTCAATCCACACCCAAA
AAGAGACCCCTCAAACCAATTTGCCTAACACGACATGGTATGGGATTTTTGGAGTAAATGTTCCCTGAAAG
TTGTATCAAGTAACATGGGTTATGAGCGATAGAGGGATCCCTAAATCTTCCGCCACATGGATGGTTTTG
CAGCCACACTTTTCAGTCTTATCAACGCTAAGGGCGAACGCTTTTGGGTGAAATTCACATTTCAACACCT
GCAAGCGTTAAGCACTTACTAACGAAAGCCGCGAGAAGTCAGAAAATATGATCCTGATTCGAATCAA
AGGGATTTATCAATGCGATCCCTAGAGGGGATTTCCCAAATGGAAATTAAGCATTCAAGTATGCCAG
AAGAAGCTTAAGAAGTATGATGATCCATCCGTTGATGTACTAATAAATTTGGTATCTCAAGATTTATCC
GTTGATGGAATGGCCATTGTAGAGTTGAATAAAAACTCTGAAAATATTTTGCAGAAGTGGAGCAAGTGC
CAATCACTCCGGCTAATGTCGTTCCGTGAATTTGGCTATAGCCCTGATAGGATGTACAAGAGCGCTTGT
TCTCTTATGGGGACACACCCGCTACCCTTAGGGGTTAATATCCTCAAATACCGGTTAATAAACCAGT
AGCCCGCTTCCACTTCTTAGCAGAGATGGTTACATGCAAAACGGTACTAGCCGCTCTTTACAAAATAT
AGCCCTAGCTCATTGCCAGGTTATAAAGAAGATAAGAGCGCGAGAGATCCTAAGTTCAACTAGCTCATA
TTGAGAAAGAGTTTGAAGTGTGAAATTTGGGATACAGGGCTGATGATAGCGATTACTCAACCCCAACAGG
TGATTACTACCGCTCATTTGCCAGCTGATGAAAAGAAGGTTGATGACACTATTTGGAGAGCTTTTGGCT
CATGTTACCCATAAGGAATTTGGGATAAACAAATTTGGAGCATTTCAAGAAGGCTGATCCCAAATACGCTG
AGGGAATTA AAAAGCTCTTGA AAAACACCAAAGATGATGAAAGACATGATGGAAAAGACATGCACCA
CACAAAAAGAAAAAGTAA
```

FASTA ▾

## Rhizoctonia solani AG1-IB WGS project CAOJ00000000 14881, whole genome shotgun sequence

GenBank: CAOJ01010216.1

[GenBank](#) [Graphics](#)

```
>CAOJ01010216.1 Rhizoctonia solani AG1-IB WGS project CAOJ00000000 data, 1
7/3/14, contig 14881, whole genome shotgun sequence
CTATAAGCACTTACCAGCCATGAATGGACAATTAACAGGGATCTGGTGAATGTTGGCACTTCCAAGACG
TGGTACTGAGCATCCGCAATAGAAGAATGCCGCGCAATGAAGCAAAGTGTCCGGAGACAGTTCAAATCCC
GGACCAAGGAACAGGAGAGACGACGCTTGTCAACATCAACGGTGAATAATCTTATCTCCACATCAGC
GAGAGCTACTCGAATAAAGGACAGAGACCTTACTTCCGATTAAGATTAATAATGATCTCCGCACTTCG
TGATCCGGAACTTGTTCGGCGCCAGCACTTGGTGAAGTCCGATGATGATGATGATGATGATGATGATG
TGAGGTTCCGGGTTGATAACTGGACAACCAAGCTCCATCGAGCGTCCGCTCCCTTTCCGATATGCTC
CCAAGTCCGGTTTGGCAAGTCAAGGTTCTTCCACACACATGGCAGACAGATTTCTCATGATGAACTCT
TTGGTCCCTTGTGGGATCGCCCAATGATACTAAAACAGCACCAATCAATCAGGTGAATACCAGAGATCAA
AGGGCTTTTTCTATAGCTCACCTTGACAAGATTCCTCTCCCTTGGCTGTTGCTCCAAAGGAAAGTATGG
CAGCCATAGCCACTCATGAAGCGCCATCCGACGGGAGTAGCGTATCCGAAAGCAGCATGAGTCCAGCGT
CTCGAGACTCGGGTACGTTTGGCAAGGAAATCAAACAGGCGTTGTAATCGATGACTCCGCTCTTGGGTT
ACGTTGCTGCGACCTGATGTTATCGGGTCCCAATAAAGGGGTCCGGAGCAAAAAGACAGGCCAGTTCAAG
CCAACCAAGTCATAGTTTCCTAGGGCAGATTTTAATAAATGATCGCAATAATGTAATGATCAAGGTATAC
CTTCTCCGTTATAAACTTAGTGGCGAACCCAGGGGATTTCCGCGCACTACAGGGAATTTCCCGTATG
GCTCCAGTCAGCAAGCAGAACGAGTGTACAATCAATCCATACTACGCGCTGACTGACAGTCCGATGAGA
GGTATAAGCAGGTTGCTCTTCCCGGTTTGGAAACAGAGCAGCCTTGGTCAGATGAGAACAATCCTTTGAA
AACTTGGAAAGTATCCAAACCCCTACAAGTCAAAAATGTTACATATTTCCGAGAAGACAGCCTTTAGACGCCAA
AACTGACCGATGATCCCGCAGCGGATGAACAATAGGTTCCGGATTTACCGCTCCGAAAGCCCTGCTGCTG
TTCTCGACACCAAGATTCCTCCCTGATACGGAATCCACCAATGTTTCAAAGCTCGTAATAAACAACAG
ATCACTACATACCAGCACTTAATACATTTCACAATCCCACTTACCAGAGTCAGGAATAGCTGTCCAAA
ATTACTAGTATAAATCGGAGATCGTCTTGGCATGAGACTCCTGAACAAGTCACGCGGAGCGAGCGGGC
GTAGCCGCGCGCCCATCAITCAATCCGCTAAACGCTTGGTTCCGCTGAATCAACCTCCGCTCCGACTCCGG
GATTTTGAAGAGCGCATCTTGAATAAACGACGCTCATGGGATTTGCAGTTTGGTTGAACCCCAAGCGGAG
GAGTAAACAGGGGACTTTGGGAGTGGGGAGGGATGAAATGTGCCCAAAGCGAGTGGTCTGACTTTTTAT
CTTTTGTGATTTGGGCTCGAAGCAATCATACGCGCATACTGGCTCATAATGACSTTTTGTCAAGT
GGTATGATTCAAATGACCAAGCGGTTTGTGAGATCGTCCGCTCCAAAGTTGGCGTGGTATGATCAAGGAT
TCAAGCGGAGAGACTGGCTTATAGAATTAAGTGGTGAAGATCAAGAACCCATGAACCCCTGAGTCTTT
```

3. Здійснюємо вхід на сторінку браузера веб-сервера EMBL-EBI – <http://www.ebi.ac.uk/services>. Вибираємо опцію «Clustal Omega» і переходимо на сторінку з вікном програми.

4. Вибираємо опцію «DNA» і в поле вставляємо сукупність нуклеотидних послідовностей. Вибираємо алгоритм вирівнювання – опція «Alignment». Більшість опцій у програмі виставлена за замовчуванням і не вимагає коректування. Після того, як усі опції встановлені натискаємо «Submit».

## 5. Отримуємо результат вирівнювання.

<https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20180>

[Input form](#) | [Web services](#) | [Help & Documentation](#)

### Results for job clustalo-I20180125-143708-0043-651

[Alignments](#) | [Result Summary](#) | [Phylogenetic Tree](#) | [Submission Details](#)

[Download Alignment File](#) | [Send to Simple\\_Phylogeny](#)

CLUSTAL O(1.2.4) multiple sequence alignment

```

CAOJ01010216.1      CTATAASCACITACCGAGCCATGAATGGACAATTAACAGGGATCTGGTGAATGTGGCAC
K02400.1             -----
U67458.1             -----

CAOJ01010216.1      TTCCAAGACGGTGGTACTGAGCATCCCGATAGAAGAATGCGGCCAATTGAAGCAAAGTGT
K02400.1             -----
U67458.1             -----

CAOJ01010216.1      CCGGAGACAGITCAATCCCGGAACCAGGGAACCAGGAGAGAACGCAGCTTGTTCACAT
K02400.1             -----AAAGGAGCAGGGCCTTTGGCTACTT
U67458.1             -----ARGAT
                        * *

CAOJ01010216.1      CACGGTGATAAATCCTTATICTCACATCAGCGAGAGCTACTCGAATAAAGGCAAGAGACCT
K02400.1      TGAGGTCACACATGACATTACCAATACTCCRAAGGCAAG-----GTATTTGAGCATATT
U67458.1      TGAGATAGAATAGTATGATAA-----TTATTAT
                        * * * * *

CAOJ01010216.1      TACTTCGGGATTACGAAITCAAATTAAGTTCTCCGACTTCGTGCA-TCGGGAACCTGTTG-
K02400.1      GGAAGAAGACTCCCATCGCAGTTCGGTTCCTCCACTGTTGC----TGGAGAATCGGGT
U67458.1      TAAAACCGAATAAAAATAAAATTTTGTITTTTAACTTTTCTTATTTTCATTAATGTTAC
                        * * * * *

CAOJ01010216.1      -CGCGCCAGACCTTGGTGACGTCGAATGCATCGAAATCG--ATTTTGGCCGAACCTGAGG
K02400.1      CAGCTGACACAGTTCGGGACCTCGTGGGTTT-GCAGTGAATTTTACACAGAAGATGGT
U67458.1      GAATAGAAATACTTAAAGGGGTTTTTCATTCTTAAAAAAGGATTTTTTAAGGAAATGA
                        * * * * *

CAOJ01010216.1      TCTTCGGGTGTGATAACCTGGACAACAACGTCATCGAGCGTCGCGCTCCCTTTTCGATA
K02400.1      AACTGGGATC-----
U67458.1      ATCTTGTAG-----
                        * *

CAOJ01010216.1      TGCTCCCACAAGTCGGTITGGCAAAGTCAGGTCITTCACCACATGCGGACAGATCTC
K02400.1             -----TCGTTGGAAT--AACACCCCAATTTCT
U67458.1             -----TCITTATATACAAAATTATGTAATATCA
                        * * * * *
    
```

---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

**1. Що є визначальним фактором при внесенні білків до групи паралогів?**

- а) належність до різних видів;
- б) належність до одного організму;
- в) рівень консервативних послідовностей;
- г) виконання різних функцій;
- д) тривимірна структура.

**2. Чи правильне твердження: *Стосовно надзвичайно подібних послідовностей між локальним та глобальним вирівнюванням немає різниці?* Дайте відповідь.**

- а) так; б) ні?

**3. Алгоритм Нідлмена-Вунша спрямований на:**

- а) локальне вирівнювання;
- б) глобальне вирівнювання.

**4. Якщо два білки мають спільного предка та виконують однакові функції у двох різних видів, їх називають .....**

- а) аналогами;
- б) ортологами;
- в) паралогами.

**5. Заповніть пропуски: *Мірою подібності двох послідовностей є* а).....?..... *та* б).....?.....**

**6. Доповніть речення: *Глобальне вирівнювання – це .....?.....***

**7. Ресурс представлений у базі UniProt, дає змогу виконати вирівнювання .....**

- а) тільки амінокислотних послідовностей;
- б) тільки нуклеотидних послідовностей;
- в) нуклеотидних і амінокислотних послідовностей.

**8. При якому вирівнюванні використовують алгоритм Сміта-Вотгермена?**

- а) локальному;
- б) глобальному;
- в) точковому.

**9. Білки, які виконують ту саму функцію в різних організмах, називаються .....**

---

---

**10. Які практичні завдання допомагає виконувати проведення вирівнювання послідовностей?**

- а) визначити їхню спорідненість;
- б) розкрити особливості тривимірної структури;
- в) встановити консервативні і варіабельні ділянки;
- г) виявити функції;
- д) з'ясувати внутрішньоклітинну локалізацію;
- е) пояснити еволюційні взаємозв'язки;

**11. Чи правильне твердження: *Мета вирівнювання послідовностей полягає в тому, щоб визначити ступінь подібності двох послідовностей і, якщо вона достатньо висока, зробити правомірний висновок про їхню гомологічність?***

- а) так; б) ні.

**12. Виберіть програми, призначені для вирівнювання послідовностей.**

- а) BLAST;
- б) DALI;
- в) Clustal;
- г) T-Coffee;
- д) RasMol;
- е) Jmol.

**13. Дайте визначення: *Відстань Левенштейна* – це .....**

**14. Які методи використовують для реалізації множинного вирівнювання?**

- а) точково-матричний;
- б) прогресивні;
- в) ітеративні;
- г) профільний аналіз;
- д) динамічне програмування;
- е) «словесні методи».

**15. Вирівнювання, яке ідентифікує схожі ділянки у межах довгих послідовностей, які дуже відрізняються на більшій частині своєї протяжності, називається:**

- а) глобальним;
- б) локальним.

**16. Доповніть речення: *Філогенетичне дерево, яке містить інформацію про довжини гілок, які представляють зміну якоїсь***

*характеристики, називається .....*

**17. Визначення спорідненості послідовностей нуклеотидів чи амінокислот встановленням відповідності мономерів двох і більше ланцюгів називається .....**

- а) ідентифікацією мішені;
- б) вирівнюванням послідовностей;
- в) встановленням кількісного співвідношення структура–активність;
- г) визначенням фармакофора.

**18. Доповніть речення: Філогенетичне дерево, яке не містить кореня і відображає зв'язок листя без передбачуваного положення загального предка, називається .....**

**19. Які із зазначених методів використовують для реалізації парного вирівнювання?**

- а) точково-матричний;
- б) прогресивні;
- в) ітеративні;
- г) профільний аналіз;
- д) динамічне програмування;
- е) «словесні методи».

**20. Вирівнювання більш ніж двох послідовностей називається .....**

**21. Заповніть пропуски: До вузла .....дерева можуть підходити чотири та більше гілок.**

**22. Із яким форматом нуклеотидних та амінокислотних послідовностей працює більшість програм із вирівнювання послідовностей?**

**23. Яке вирівнювання подане на рисунку?**

CLUSTAL 2.1 multiple sequence alignment

```
***_S30M_2508      -----ATG-----AGTAGT-----TTGTGTGTGTTAA-CGATA----- 28
abc_S60S012_E2191 -----ATG-----AGTAGT-----TTGTGTGTGTTAA-CGATA----- 28
sec_BC2433         -----ATGGGACAGAGAGATGATATTCTGTCAATGCSCTTCCA-CGACA----- 43
***_BTM2405       -----ATGGGACAGAGAGATGATATTCTGTCAATGCSCTTCCA-CGACA----- 43
bta_D14105        -----ATG-----AG-----GA-GGAGG----- 10
hwa_B166          -----ATG-----AG-----GA-GGAGG----- 10
lpe_DP1821        -----ATG-----ATGAA-----GCTGGCGCCGCCATCCGCA----- 29
xoo_X002033       TTGCGCATCTTCCCAACCGGCGCCGCGCATGAGATGCAACCGACAGCCACTCGTC 60
hru_BF028020     -----ATG-----TCATATCGATA----- 13
btp_DP_4435      -----ATG-----TATGCAATTAAG----- 16
**
```

- а) парне;
- б) множинне.

---

---

**24. Яке вирівнювання призначене для порівняння послідовностей, у яких збігаються тільки кінцеві ділянки?**

**25. Під час якого методу множинного вирівнювання спочатку порівнюють найбільш схожі послідовності і поступово додають менш споріднені?**

- а) динамічного програмування;
- б) прогресивних методів;
- в) ітеративних методів;
- г) профільного аналізу.

**26. Чи дає змогу програма BLAST провести вирівнювання нуклеотидних послідовностей?**

- а) так; б) ні.

**27. Дайте визначення: *Кладограма* – це .....**

**28. Філогенетичне дерево, у якому до кожного вузла підходять рівно три гілки, називається .....**

- а) біфуркаційним;
- б) мультифуркаційним.

**29. Основне призначення програми ClustalW .....**

- а) здійснення множинного вирівнювання;
- б) визначення тривимірної структури білків та нуклеїнових кислот;
- в) обчислення еволюційних дистанцій між послідовностями;
- г) передбачення вторинної структури макромолекул;
- д) визначення характеру та типу амінокислотних замінів.

**30. Чи правильне твердження: *Марковане філогенетичне дерево містить назви листків, натомість немарковане просто відображає топологію:***

- а) так; б) ні.



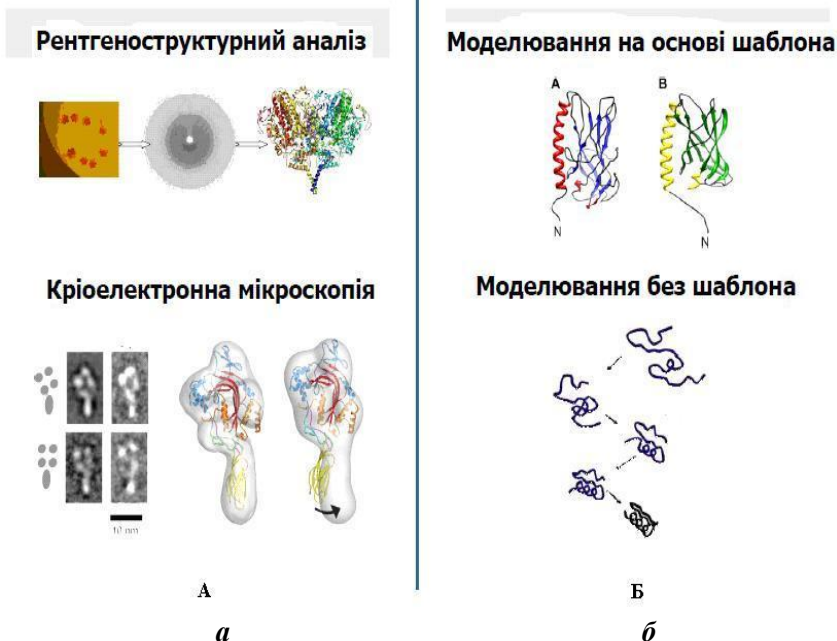
## ТЕМА 4. МЕТОДИ ПЕРЕДБАЧЕННЯ СТРУКТУРИ ТА ФУНКЦІЙ БІЛКІВ

**Мета:** ознайомитися із методами структурного вирівнювання протеїнів. Навчитися використовувати бази даних та різноманітні програми для моделювання вторинної та третинної структур білків та передбачення їхніх функцій.

### ТЕОРЕТИЧНА ЧАСТИНА

#### 1. Складність проблеми і типи методів моделювання білків

Отримати 3D-структури білків сьогодні можна завдяки двом підходам – експериментальним (рентгеноструктурний аналіз, ЯМР-спектроскопія, криоелектронна мікроскопія) або теоретичним (моделювання тривимірної структури білка на основі шаблону або без шаблону).



*Рис. 23. Експериментальні (а) та теоретичні (б) підходи моделювання 3D-структури білка*

---

---

Більшість просторових структур білків і нуклеїнових кислот отримані саме методом рентгенівської кристалографії. Метод полягає в тому, що кристали білка опромінюють. Детектор променів передається на комп'ютер та будується карта електронної густини, а в подальшому – просторова модель білка.

ЯМР-спектроскопія допомагає визначати структуру молекул у розчині (без кристалів) та оцінювати динаміку молекул.

Метод кріоелектронної мікроскопії має низьку роздільну здатність, але дає змогу розкрити структуру макромолекулярних комплексів.

Комп'ютерне моделювання третинної структури білка базується на амінокислотній послідовності і сприяє передбаченню його просторової структури.

Моделювання тривимірної структури білка – одна з найважливіших проблем у сучасній біоінформатиці. Якщо публічні бази даних сьогодні містять понад 10 млн послідовностей амінокислот білків, то централізовану базу даних структур білків PDB вдалося наповнити лише до 60 тисяч структур. Це пов'язано насамперед із тим, що методи експериментального визначення структури білка технологічно складні, коштовні і дуже відстають у продуктивності від методів визначення хімічного складу. Тому заповнити прогалину між кількістю послідовностей і 3D-структур білків можна лише теоретичним моделюванням структури білків.

Головна проблема в моделюванні білків, безумовно, не побудова самої тривимірної моделі, а визначення дійсної конформації досліджуваного білка. Річ у тому, що в 1968 році Сайрус Ловінталь обчислив і зауважив, що для білка, який має 100 амінокислотних залишків, є приблизно  $100^{100}$  конформацій. Причому за частки секунди амінокислотний ланцюг формує найбільш енергетично вигідний варіант. Дане явище дотепер не має пояснення, що є основним каменем спотикання в моделюванні білкових молекул. Саме тому моделювання виконують на основі гомологів із бази даних, 3D-структури яких виявлені емпірично.

При передбаченні структури білків дослідник стикається з двома проблемами:

---

1) кількість можливих просторових конфігурацій білків дуже велика;

2) фізичні засади структуроутворення білків і їхньої стабільності ще не до кінця зрозумілі.

На даний час існують два головні, концептуально різні типи методів для моделювання структури білків: інформаційні методи та фізичні методи.

Інформаційні методи використовують припущення, що невідома структура білка може бути схожою до однієї або кількох відомих структур білків, чи, принаймні, бути складеною з елементарних конструкційних блоків таких білків. Цей тип методів ще називають моделюванням за зразком. Шаблони можуть бути знайдені за допомогою методів безпосереднього порівняння амінокислотних послідовностей. Моделювання за шаблоном має величезний практичний потенціал, адже якщо відома структура хоча б одного білка з якоїсь функціональної родини, то тоді можна спробувати побудувати моделі для практично кожного білка в цій родині. Із збільшенням бази даних структур таке моделювання стає можливим для все більшої кількості білків.

Коли не вдається знайти шаблон для моделювання, тоді застосовуються так звані безшаблонні методи до яких належать фізичні методи, які враховують деталі взаємодій на атомному рівні.

Отже, одним із завдань біоінформатики є встановлення функціональної залежності між послідовністю амінокислот і просторовою структурою білка. Якщо така залежність буде встановлена, то можна буде достатньо точно передбачити структуру білка за послідовністю амінокислот.

## **2. Передбачення вторинної структури білка**

Одне із важливих завдань аналізу послідовностей – точне передбачення джерел формування  $\alpha$ -спіралей,  $\beta$ -листів та інших елементів вторинної структури в амінокислотному ланцюгу білка.

Якість передбачення вторинної структури залежить від точності розпізнавання типу елементів вторинної структури, а також від точності визначення місцезнаходження та довжини

---

цих елементів. Основними типами вторинної структури, які досліджуються, є  $\alpha$ -спіралі,  $\beta$ -листи і супервторинні структури.

Фізико-хімічний процес, у результаті якого білки в своєму природному середовищі (розчині, цитоплазмі або мембрані) набувають характерного лише для них просторового укладання та функцій, називається *фолдингом*. Іншими словами, фолдинг – це згортання білків (та інших біомакромолекул) із розгорнутої конформації на «нативну» форму.

Термін «ab initio» (з перших принципів)-фолдинг часто вживається для позначення методів комп'ютерного передбачення структури білка без використання структурних даних про інші білки. У передбаченні структури білків не використовують у явній формі інформацію про структуру інших білків. Всі обчислення, як правило, здійснюються в емпіричних силових полях, які описують парні взаємодії в класичній системі часток, котрі представляють молекулу білка. Самі ж ці силові поля в неявній формі містять дані про структуру молекул – парціальні заряди та масу атомів, а також довжини і кути валентних зв'язків. Тому інколи застосовують термін «de novo»-фолдинг (заново).

Нині найширше використання в передбаченні вторинної структури білків мають такі методи.

#### 1. Метод Чоу-Фасмена.

Метод, запропонований Пітером Чоу та Джеральдом Фасменом, оснований на тому, що кожна амінокислота індивідуально впливає на вторинну структуру в межах певних послідовностей. Ґрунтується на аналізі частоти появи кожної з 20 амінокислот в  $\alpha$ -спіралях,  $\beta$ -листах і  $\beta$ -згинах. Окрім того, побудований на ньому алгоритм використовує спеціальний набір правил передбачення вторинної структури. Алгоритм, розглядаючи послідовність, намагається знайти коротку підпослідовність амінокислот, яка показує високу схильність до певного структурного типу білка.

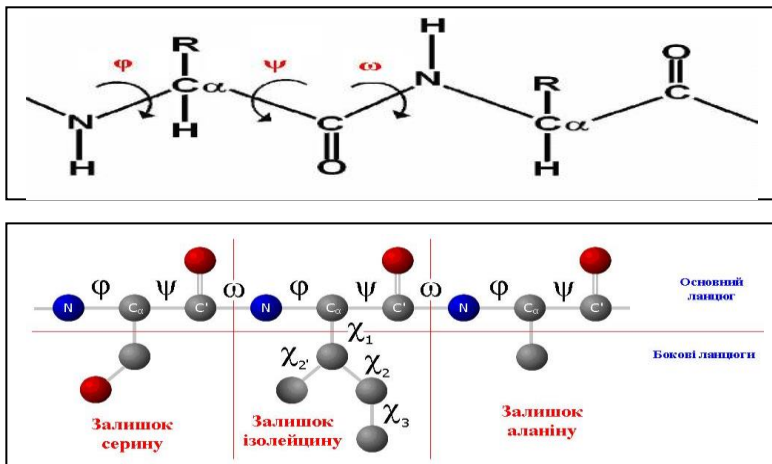
#### 2. Методи моделювання нейронних мереж.

У підході нейронних мереж комп'ютерні програми розпізнають регуляторні комбінації амінокислот, розміщені у відомих вторинних структурах і відрізняють ці комбінації від інших амінокислотних груп, не наявних у цих структурах.

### 3. Методи пошуку «найближчого сусіда».

Метод передбачає розміщення амінокислот з послідовності у визначеній конформації вторинної структури.

В основі передбачення вторинної структури білків лежать торсійні кути поліпептидного ланцюга. В поліпептидному ланцюзі є три торсійні кути –  $\phi$  (фі),  $\psi$  (псі) і  $\omega$  (омега):

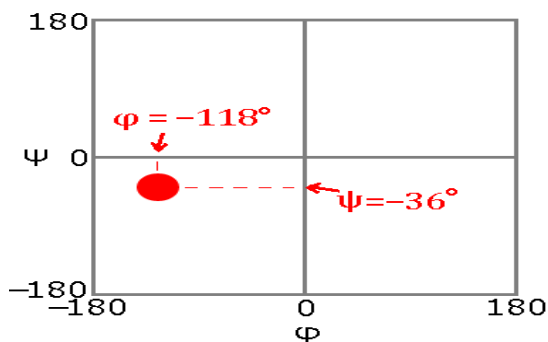


*Рис. 24. Торсійні кути в поліпептидних ланцюгах*

Обертання навколо  $\omega$ -кута надзвичайно обмежене. Обертання кутів  $\psi$  і  $\phi$  вільне і вони теоретично можуть набувати будь-які значення. Саме ця обставина зумовлює наявність різноманітних просторових структур білків, оскільки на кожен амінокислотний залишок припадає по 2 міри свободи (виняток – пролін, у якого лише 1 міра свободи). На практиці в структурах білків часто виявляються локальні обмеження в конкретних ділянках білкового ланцюга для обертання торсійних кутів, що зумовлене як стеричними чинниками (контакти бічних ланцюгів близько розміщених амінокислот у ланцюзі), так і взаємодіями амінокислотних залишків при формуванні елементів вторинних структур (взаємодія амінокислот, які далеко розміщені в послідовності, але зближуються у просторі).

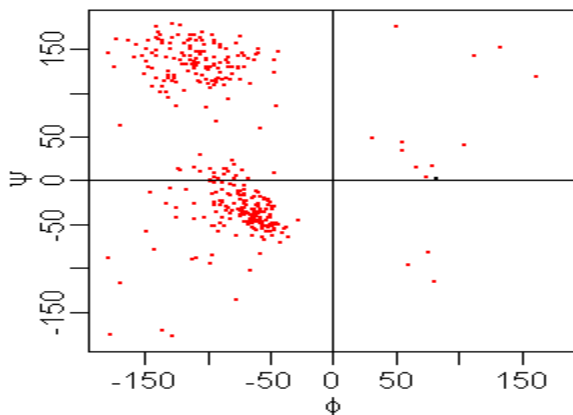
Для опису конформації цілих молекул білка широко використовують карти Рамачандрана.

Згідно з побудовою цієї карти, конформацію білкового ланцюга можна подати у вигляді набору точок на двовимірній діаграмі, де кожна точка відповідатиме одному амінокислотному залишку в структурі білка, а значення двох торсійних кутів і цього залишку є координатами на площині. Положення точки по горизонталі показує кут  $\phi$ , по вертикалі –  $\psi$ .



*Рис. 25. Положення амінокислоти на карті Рамачандрана*

Така діаграма досить зручний засіб формального опису та конформаційного аналізу просторової структури білків. Карта Рамачандрана (Ramachandran Plot) для білків із великою кількістю амінокислот матиме такий вигляд (рис. 26):

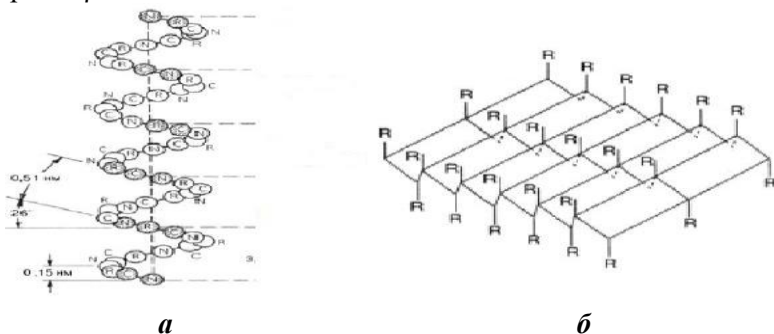


*Рис. 26. Карта Рамачандрана для фосфорибозилтрансферази*

Кожна точка на карті позначає один амінокислотний залишок.

Зауважимо, що карта Рамачандрана не визначає третинну структуру білка однозначно, оскільки не показує порядок поєднання амінокислот. Вона лише допомагає побачити переважну конформацію амінокислот у білку. На практиці карти Рамачандрана використовуються для демонстрації наявності в білку тих або інших вторинних і надвторинних структур.

Кarti Рамачандрана допомагають виявити так звані регулярні структури – ділянки поліпептидного ланцюга, в яких амінокислотні залишки мають схожу конформацію (в тому сенсі, що у них приблизно рівні кути  $\phi$  і  $\psi$ ). Конформація радикалів не розглядається. Регулярні структури – різновид вторинних структур. Прикладами регулярних структур є  $\alpha$ -спіралі і  $\beta$ -листи.



**Рис. 27. Вторинні структури білка**

*Примітка: а –  $\alpha$ -спіраль; б –  $\beta$ -лист.*

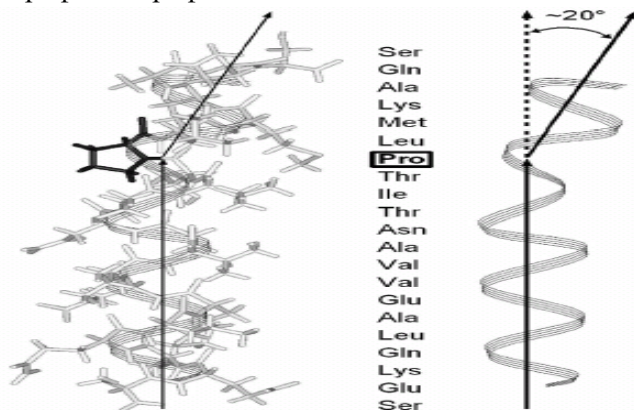
Якщо спробувати визначити кути  $\phi$  і  $\psi$  амінокислотних залишків  $\alpha$ -спіралі, можна помітити, що вони приблизно однакові у різних залишках і мають значення в межах від  $-40$  до  $-60$ . Відповідно, на карті Рамачандрана для даної структури точки концентруватимуться у певній ділянці. Значна кількість точок у ділянці, характерній для тієї або іншої вторинної структури, дає змогу судити про наявність цієї структури в білку.

Стандартні значення кутів для вторинних структур у білках такі:

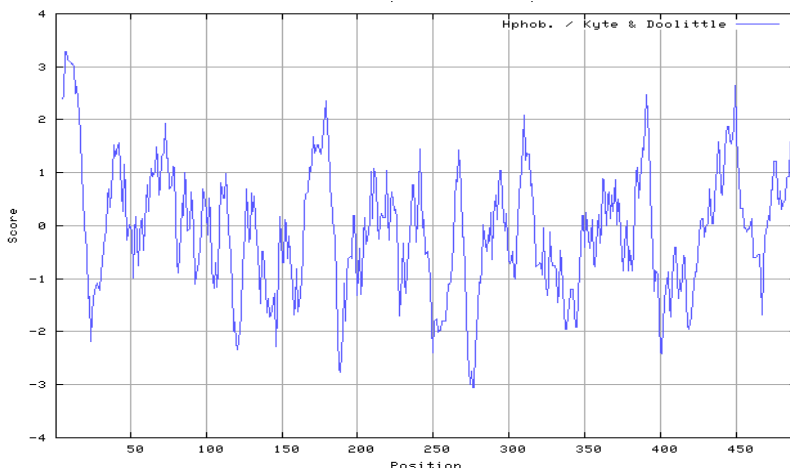
правозакручена  $\alpha$ -спіраль –  $\phi = -57^\circ$ ,  $\psi = -47^\circ$ ;

паралельний  $\beta$ -лист –  $\phi = -119^\circ$ ,  $\psi = +113^\circ$ ;  
 антипаралельний  $\beta$ -лист –  $\phi = -139^\circ$ ,  $\psi = +136^\circ$ ;  
 лівозакручена  $\alpha$ -спіраль —  $\phi = +57^\circ$ ,  $\psi = +47^\circ$ ;  
 поліпролінова спіраль –  $\phi = -79^\circ$ ,  $\psi = +150^\circ$ .

Для встановлення згинів у вторинній структурі білків будують профіль гідрофобності.



**Рис. 28.** Схема зображення згину у вторинній структурі білка



**Рис. 29.** Профіль гідрофобності

Значення нижче -2 вказують на згини у білковій молекулі.



При моделюванні вторинної структури білка використовують такі поняття, як конформер та ротамер.

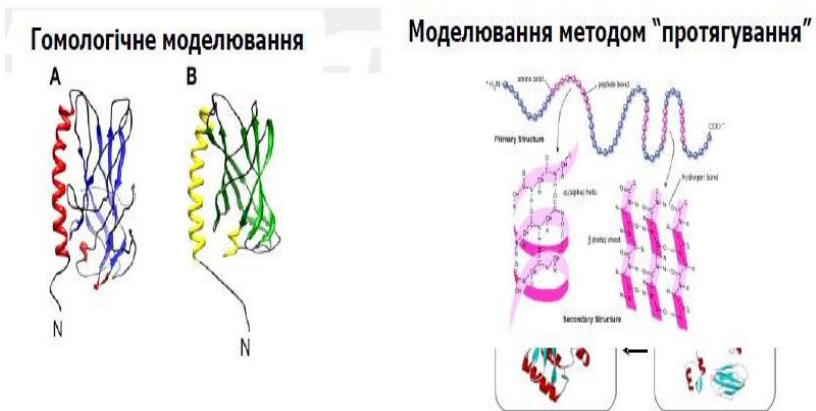
*Конформер* – особливий згорнутий стан або структура білка, яка має низьку енергію, тобто є термодинамічно стійкою при фізіологічних умовах.

*Ротамер* – один із конформерів, який є результатом обмеженого обертання навколо єдиного хімічного зв'язку.

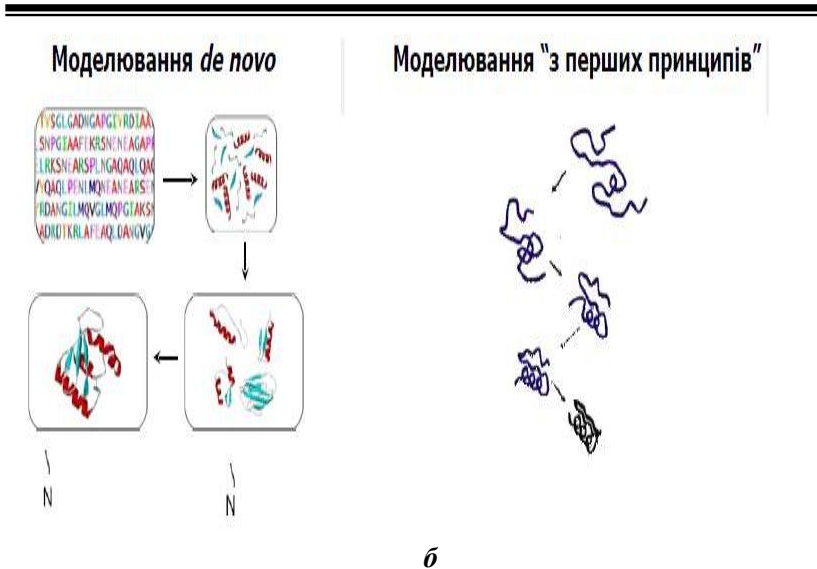
Завдяки великому обсягу баз даних створені бібліотеки ротамерів, які використовуються в молекулярному моделюванні: алгоритм виділяє з них найімовірніші конформації бокових ланцюгів амінокислот і додає до основного ланцюга макромолекули.

### 3. Моделювання просторової структури білка

Як уже зазначалося методи моделювання структури білків поєднуються у два типи – моделювання на основі шаблону (до цих методів належить гомологічне моделювання та моделювання методом «протягування») та моделювання без шаблону (моделювання «de novo» та «з перших принципів»).



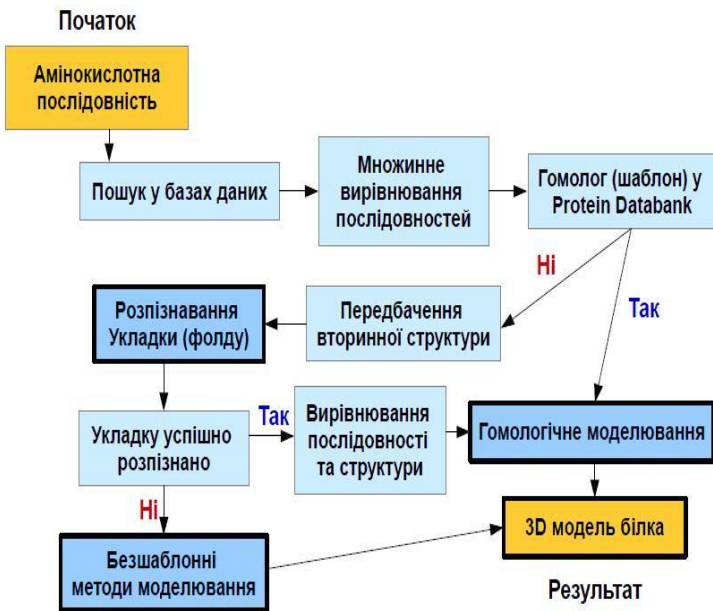
*a*



**Рис. 30. Методи моделювання структури білків на основі шаблону (а) та без шаблону (б)**

Є загальна стратегія вибору методу моделювання. Наприклад, нас цікавить білок, структуру якого досі не визначено експериментально. Як обрати адекватний метод моделювання?

Спочатку знаходимо амінокислотну послідовність у базах даних. Після множинного вирівнювання у базі даних PDB шукаємо шаблон-гомолог. Якщо гомолог знайдено, виконуємо гомологічне моделювання, якщо ні – передбачаємо вторинну структуру білка. Після передбачення вторинної структури розпізнаємо укладку (фолд). І тут знову є два шляхи: якщо укладку знайдено, здійснюють вирівнювання послідовності і структури, а далі гомологічне моделювання; якщо ні – використовуємо безшаблонні методи моделювання (рис. 31).



*Рис. 31. Алгоритм вибору методу моделювання білків*

### **Гомологічне моделювання**

Суть методу полягає у побудові тривимірної моделі білка, структура якого невідома, на основі подібності його амінокислотної послідовності білка, структуру якого встановлено експериментально.

Метод базується на двох положеннях:

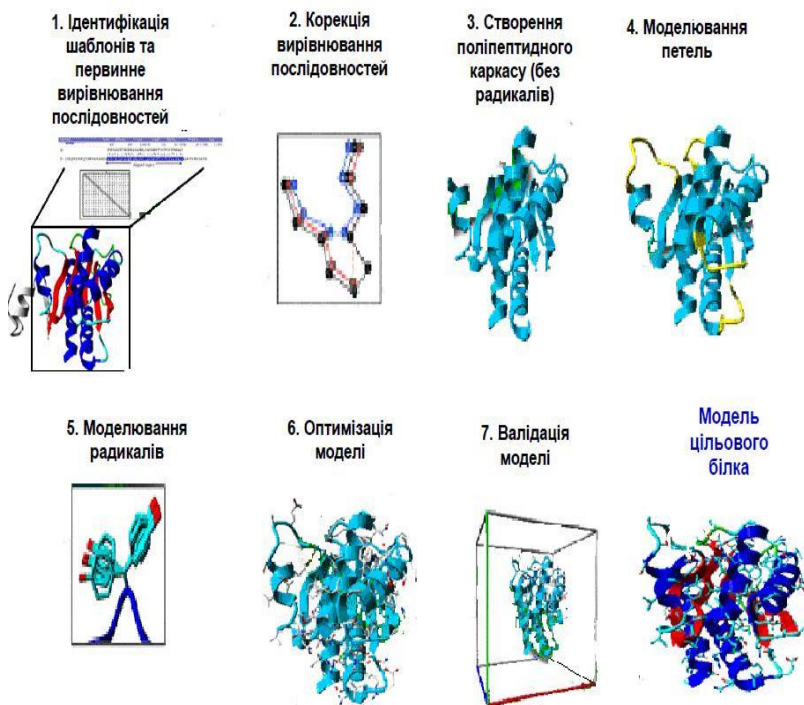
- 1) структура білка повністю визначається його амінокислотною послідовністю;
- 2) структура більш консервативна, ніж послідовність.

Це найефективніший метод моделювання білків. Чим вища ідентичність послідовностей цільового білка та білка-шаблону, тим надійніше моделювання та менша роль людини, і навпаки. Як-от, якщо подібність послідовностей більше ніж 75 %, можливе повністю автоматичне моделювання, створюються високоякісні моделі; 50–75 % – якість моделей залежить від ефективності оптимізації людиною; 25–50 % – якість залежить

від вирівнювання (можливе створення як повністю помилкових, так і дуже якісних моделей); менше 25 % – основна проблема – пошук правильного шаблону, потрібні додаткові підходи та досвід.

Можна виділити сім основних етапів гомологічного моделювання для створення моделі цільового білка:

1. Ідентифікація шаблонів та первинне вирівнювання послідовностей.
2. Корекція вирівнювання послідовностей.
3. Створення поліпептидного каркасу (без радикалів).
4. Моделювання петель.
5. Моделювання радикалів.
6. Оптимізація моделі.
7. Валідація моделі.

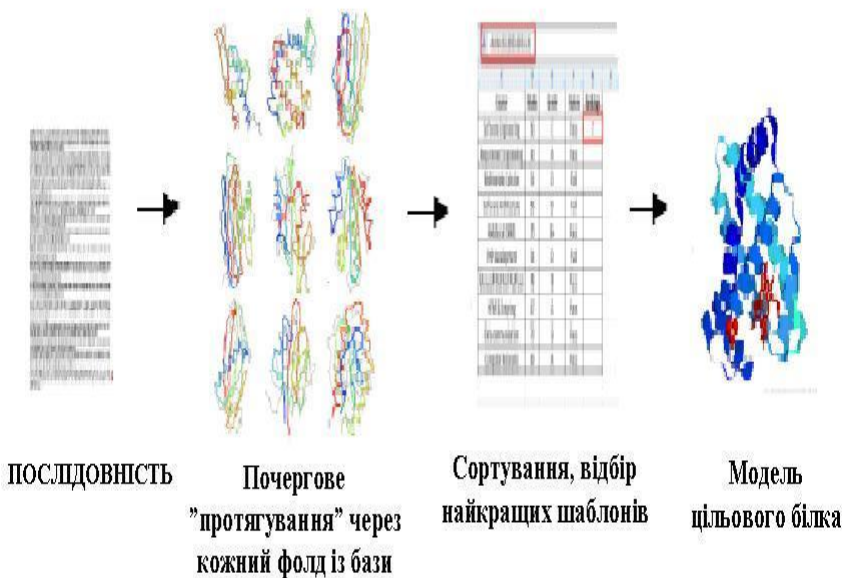


*Рис. 32. Основні етапи гомологічного моделювання*

### **Метод протягування (розпізнавання укладки)**

За приблизними оцінками, є тільки близько 2000 білкових фолдів (укладок), натомість білкових послідовностей – десятки мільйонів. Це означає, що багато послідовностей без явної гомології теж можуть мати подібну просторову укладку.

Метод використовує систему оцінювання сумісності послідовності-мішені з кожним із фолдів у бібліотеці баз даних. Послідовність білка-мішені ніби почергово «протягується» через кожний фолд. Далі відбувається сортування та відбір найкращих шаблонів. У кінцевому результаті створюється модель цільового білка.



**Рис. 33. Основні етапи моделювання білків методом «протягування»**

Спеціальні бібліотеки фолдів можна знайти в базах даних CATH, SCOP, RaptorX, FUGUE, Phyre2.

### *Моделювання білків de novo*

Метод має проміжне положення між повністю безшаблонним передбаченням структури білка та гомологічним моделюванням. У літературі часто поєднується із *ab initio* моделюванням, хоча це різні методи.

Знаючи послідовність, на першому етапі здійснюють випадкове збирання фрагментів та створюють моделі низької якості. На другому – фізичну оптимізацію моделі, внаслідок чого отримують модель цільового білка.

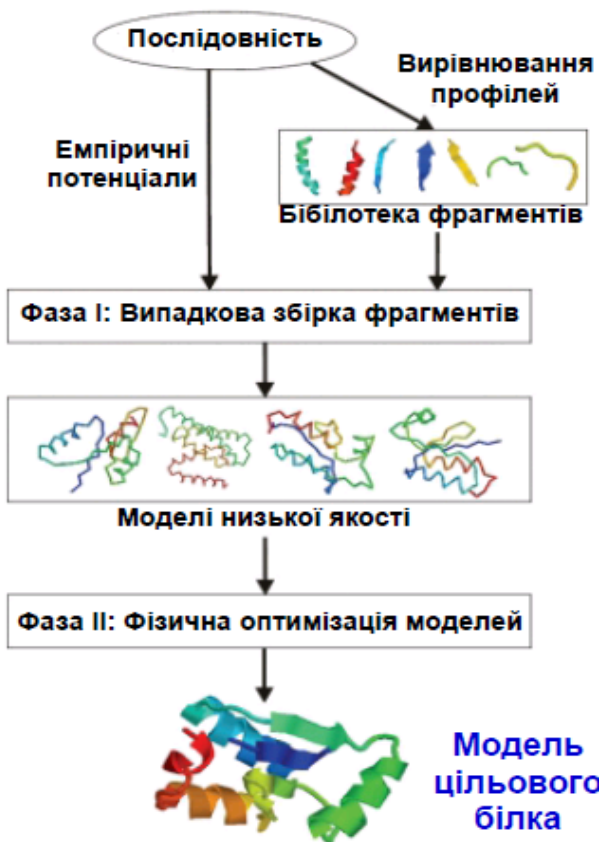


Рис. 34. Основні етапи моделювання білків *de novo*

### **Моделювання «з перших принципів» (*ab initio*)**

Якщо придатних шаблонів для моделювання немає, можна скористатися методами вільного моделювання (передбачення «з перших принципів»), що повністю базуються на фізичних принципах.

Цей метод ґрунтується на припущенні того, що найнижче значення вільної енергії відповідає нативній структурі. При даному виді моделювання виконують молекулярно-механічну симуляцію білка на основі первинної структури. Далі відбирають моделі з мінімальною енергією та створюють модель цільового білка.

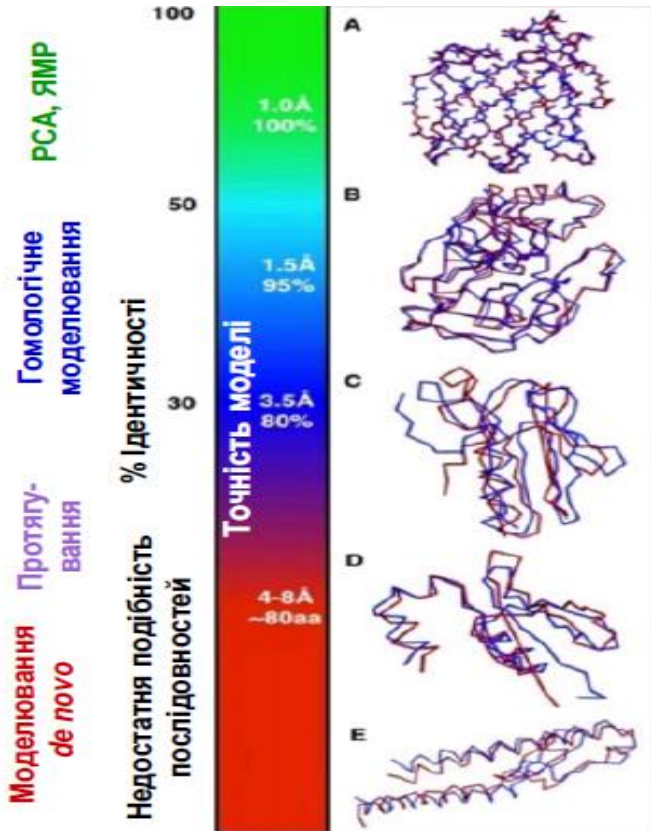


**Рис. 35. Основні етапи моделювання білків методом *ab initio***

Основні недоліки моделювання ab initio:

- ефективні лише для поліпептидів, які містять менш як 100 амінокислотних залишків;
- надзвичайно вимогливі до обчислювальних ресурсів (потребують суперкомп'ютерів).

Якщо порівняти охарактеризовані методи моделювання, то найефективнішим є гомологічне моделювання.

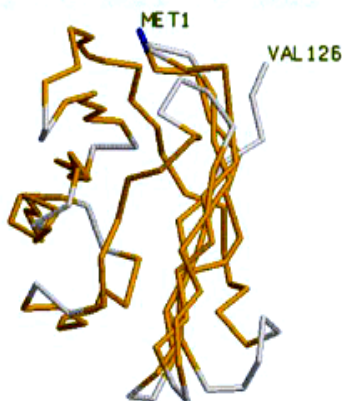


*Рис. 36. Ефективність методів передбачення третинної структури білка*

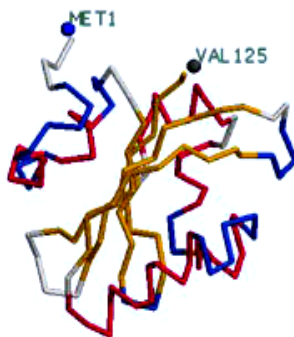


---

Часом у PDB трапляються повністю помилкові моделі. Так, якщо подивитися на дві моделі фотоактивного жовтого білка рецептора фототаксису *Ectothiorhodospira halophita*, встановлені з інтервалом 6 років, то важко побачити щось спільне.



1PHY, 2.4 Å  
McRee et al. 1989



2PHY, 1.4 Å  
Borgstahl et al. 1995

Отже, основа функціональності білка, яка вимагає точної просторової організації великої кількості амінокислот, – третинна структура. Організація білкових структур відповідно до їхніх патернів укладки дуже зручна для виявлення даних у базі PDB. На цьому оснований принцип пошуку інформації. Кілька баз даних – похідних від PDB – побудовані на класифікації білкових структур. У них пропонуються зручні інструменти для вивчення структур білків, як-от пошук білка за ключовим словом і послідовностями, навігація серед подібних структур на різних рівнях систематичної ієрархії, сканування бази даних на предмет пошуку структур, подібних із заданою структурою і посилання на інші сайти. Прикладами таких баз даних є SCOP і CATH, які використовують для пошуку гомології в білкових родинах.

---

---

#### 4. Програми структурного вирівнювання білків. Передбачення функцій білків

Оскільки структура протеїнів консервативніша, ніж послідовність, структурне вирівнювання допомагає точніше визначати функціональну й еволюційну спорідненість протеїнів. Окрім того, структурне вирівнювання – досить надійний спосіб для передбачення функції протеїнів. Для структурного вирівнювання потрібна інформація щодо вторинної і третинної структур протеїнів. Отже, очевидно те, що ці методи застосовуються лише до послідовностей із експериментально встановленою просторовою структурою. Результати структурного вирівнювання можуть також бути основою для гомологічного моделювання та структурної класифікації.

Методи структурного вирівнювання використовують для порівняння двох або більшої кількості послідовностей і для здійснення локальних вирівнювань.

Програма *DALI* ([http://ekhidna.biocenter.helsinki.fi/dali\\_server](http://ekhidna.biocenter.helsinki.fi/dali_server)) призначена для попарного структурного вирівнювання протеїнових структур. За допомогою просторових координат кожного протеїну обчислюють матриці відстаней між  $C\alpha$ -атомами амінокислотних залишків. Матриці відстаней спочатку розкладають на прості контактні мотиви, наприклад на субматриці гексапептид-гексапептид. Потім подібні контактні мотиви у двох матрицях розміщують попарно і комбінують у більш послідовні ряди пар.

Для конструювання бази даних структурного вирівнювання використаний метод DALI. На веб-сервері DALI можна порівняти структури досліджуваного протеїну з наявними у PDB.

*RasMol* (доступ на сайті <http://www.openrasmol.org>) – комп'ютерна програма, за допомогою якої візуалізуються молекули. Використовується вона переважно для вивчення й отримання зображень просторових структур біологічних макромолекул, насамперед білків і нуклеїнових кислот. Перша версія програми RasMol створена Роджером Сейлом на початку 1990-х років. Вихідними даними для візуалізації є координати атомів молекули (або комплексу молекул), які містяться у файлі формату PDB.

---

---

При візуалізації білків у вікні програми зображаються різні моделі. Найбільш уживані моделі такі:

дротяна модель – ковалентні зв'язки між атомами зображаються лініями, які з'єднують їхні центри. RasMol, як правило, визначає наявність ковалентних зв'язків за відстанню між центрами атомів;

кулькова – атоми зображаються кульками. Накладання кулькової та дротяної моделей іноді називають шарнірною моделлю;

остовна модель – зображаються умовні лінії, які з'єднують Ca-атоми. Робота відбувається у двох вікнах – графічному та командному.

**Jmol** (<http://jmol.sourceforge.net>) – програма для перегляду структури молекул у трьох вимірах. Jmol використовується як для навчальної мети, так і під час виконання наукових досліджень у сфері молекулярної біології, хімії і біохімії. Програма вільна і відкрита. Вона дає змогу будувати зображення молекул у різний спосіб. Jmol підтримує велику кількість форматів файлів.

**BioEditor** (<http://en.bio-soft.net/3d/bioeditor.html>) – програма для опису структури макромолекул. Допомогає створювати опис, який містить форматований текст, графіку, сіквенси та інтерактивні зображення молекул.

**Modeller** (<https://salilab.org/modeller>) – програма для гомологічного або порівняльного моделювання тривимірної структури білка. Вона в автоматичному режимі розраховує модель білка за поданими вирівнюваннями послідовностей модельованих білків із відомими аналогами і допомагає здійснювати подальший аналіз структури.

**LINUS** (<http://roselab.jhu.edu/dist/manual/index.html>) – програма передбачення структури білка за амінокислотною послідовністю. Процедура повністю апріорна, тобто працює тільки із самою послідовністю, не спираючись ні на експериментальні дані, ні на відомі структурні кореляції. У LINUS реалізований ієрархічний алгоритм – згортання починається з коротких фрагментів, поступово об'єднуючи їх у довші.

---

---

***Ramachandran Plot Explorer*** (<http://boscoh.com/ramaplot>) – крихітна програма не потребує інсталяції для візуалізації та конформаційного аналізу білків і поліпептидів. Створена тільки для роботи з  $\phi$ ,  $\psi$  і  $\chi$  кутами і дає змогу:

- аналізувати конформаційні зміни;
- аналізувати вторинну структуру і загальну структуру петель;
- інтерактивно конструювати та редагувати білковий сіквенс.

Багато програм передбачення просторових структур невідомих білків за наявністю відомих хімічних і фізичних властивостей амінокислот доступні через сервери: ExPASy, EMBL, EBI, The Wellcome Trust Sanger Institute (<http://www.sanger.ac.uk/resources/software>), університету Манчестера (<http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/PRINTS.html>), The Institute for the Biology and Chemistry of Proteins (IBCP) ([http://pbil.ibcp.fr/htm/index.php?page=pbil\\_ibcp\\_Software.html](http://pbil.ibcp.fr/htm/index.php?page=pbil_ibcp_Software.html)).

### ***Передбачення функцій білків***

Наявність подібних просторових структур протеїнів часто свідчить про схожі функції, але є випадки, коли протеїни з однаковою укладкою виконують різну біохімічну роль. Протеїни з різними укладками можуть мати однакові функції. Найбільш вірогідним є припущення, що такі зв'язувальні мотиви визначають схожі функції протеїнів.

У процесі еволюції білки можуть:

- зберігати функцію і специфічність;
- зберігати функцію, але змінювати специфічність;
- змінювати функцію на подібну або ту ж саму, але в іншому метаболічному «контексті»;
- переключатися на зовсім іншу функцію.

Часто зміна функції зовсім не пов'язана зі зміною білкової послідовності або структури.

Спільність або відмінність функцій білків нерідко залежить від функціональної дивергенції або конвергенції білків.

*Функціональна дивергенція* – процес, за допомогою якого білки, після дуплікації генів, набувають функцій, відмінних від предкового білка. Функціональна дивергенція може привести або до субфункціональності, де паралоги виконують одну фун-

---

---

кцію з кількох подібних, або нефункціональності, коли утворений білок виконує абсолютно нову функцію.

Вважається, що процес дуплікації генів і функціональної дивергенції білків – одна з причин утворення великої групи білкових родин, наявних на сьогодні.

Дуплікація генів – чинник функціональної дивергенції білків. Функціональна розбіжність серед білків можлива також унаслідок мутацій або горизонтального перенесення генів.

Приклад функціональної дивергенції білків – розбіжності гемоглобіну та міоглобіну, які виконують в організмі різні функції: гемоглобін транспортує кисень від легень до тканин; міоглобін – кисень-зв'язувальний білок скелетних м'язів та м'язів серця хребетних тварин (тобто дані білки – паралоги).

*Функціональна конвергенція білків* – це процес, який зумовлює формування комплексу схожих ознак у представників неспоріднених груп білків.

Так, гексокіназа, рибокіназа і галактокіназа проявляють подібні ферментативні функції – фосфорилування цукрів, але вони розвивалися з трьох різних негомологічних родин, мають різну тривимірну структуру і їхні послідовності суттєво відрізняються.

Функціональну конвергенцію білків можна прослідкувати на прикладі антифризних білків, яким властива здатність знижувати точку замерзання води, прямо зв'язуючись з поверхнею льодових кристалів, порушуючи їхню структуру. Антифризні білки знайдені в деяких риб, комах і рослин, які живуть у холодних умовах. За структурою антифризні білки різних організмів суттєво відрізняються, однак виконують однакові функції.

Гемоціанін членистоногих і моллюсків еволюціонував від різних предків. Так, гемоціанін членистоногих утворився з тирозинази, гемоціанін моллюсків – із протеїнів комах. Гемоціаніни цих видів мають різну молекулярну масу та структуру. Проте обидва білки містять мідь та беруть участь у зв'язуванні та транспортуванні кисню.

Отже, патерн фолдингу не завжди може точно вказувати на функцію білка, особливо у випадку віддалених гомологів. Тому основна увага при передбаченні функцій повинна приділятися

---

---

активному центру. Ідентифікацію подібних функціональних сайтів застосовують для дизайну ліків, зокрема виявлення протеїнів, на які може впливати досліджувана сполука і такий спосіб призводить до побічних ефектів.

Вклад біоінформатики в передбачення функції білка за його амінокислотною послідовністю і структурою, ймовірно, не буде вичерпуватися одним лише алгоритмом, який встановлює однозначний зв'язок між функцією та структурою.

## ПРАКТИКУМ

### Завдання 1

Використовуючи Інтернет-ресурси дослідіть параметри вторинної структури для лізоциму людини. Побудуйте профіль гідрофобності для даного білка, передбачте найменшу кількість згинів (показник -2) та порівняйте її з даними інших сайтів.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс UniProt (<http://www.uniprot.org>). Вводимо запит у пошуковому полі UniProt для лізоциму людини (human lysozyme) і натискаємо «Search». З результатів пошуку вибираємо той, який відповідає лізоциму людини (наприклад, з ID P61626).

The screenshot shows the UniProtKB search results for 'human lysozyme'. The search bar contains 'human lysozyme'. Below the search bar, there is a yellow notification box about the transition from HTTP to HTTPS. The main content area is titled 'UniProtKB results' and shows a table of search results. The table has columns for Entry, Entry name, Protein names, Gene names, and Organism. The first entry is P61626, LYSC\_HUMAN, Lysozyme C, with gene names LYZ and LYZLZM, and organism Homo sapiens (Human). Below this, there is a section for 'Associated proteins' showing a table with columns for Entry, Entry name, Protein names, Gene names, and Organism. The first entry in this table is P00698, LYSC\_CHICK, Lysozyme C, with gene name LYZ and organism Gallus gallus (Chicken). Other entries include P42684, ABL2\_HUMAN, Abelson tyrosine-protein kinase 2, with gene names ABL2, ABL, ARG and organism Homo sapiens (Human); P00709, LALBA\_HUMAN, Alpha-lactalbumin, with gene names LALBA, LYZL7 and organism Homo sapiens (Human); P02489, CRYAA\_HUMAN, Alpha-crystallin A chain, with gene names CRYAA, CRYA1, HSPB4 and organism Homo sapiens (Human); O75951, LYZL6\_HUMAN, Lysozyme-like protein 6, with gene names LYZL6, LYC1, UNQ754/PRO1485 and organism Homo sapiens (Human); and Q96K00, LYZL4\_HUMAN, Lysozyme-like protein 4, with gene names LYZL4, LYC4 and organism Homo sapiens (Human).

2. Переходимо на сторінку результату та знаходимо інформацію про вторинну структуру білка, яка зображена у вигляді різнокольорової стрічки. Синім кольором позначені  $\alpha$ -спіралі, зеленим –  $\beta$ -листи, червоним – згини, сірим – невпорядковані ділянки. Аналіз результату пошуку показав, що білок містить 5 згинів.

www.uniprot.org/uniprot/P61626

GO - Molecular function

- identical protein binding Source: IntAct

View the complete GO annotation on QuickGO ...

Protein-protein interaction databases

BioGRID	110247: 32 interactors.
IntAct	P61626: 10 interactors.
MINT	MINT-5002660.
STRING	9506.ENSP00000261267.

Structure

Secondary structure

1

Legend: Helix Turn Beta strand | PDB Structures known for this area

Show more details

3D structure databases

3. Для побудови профілю гідрофобності здійснюємо вхід на веб-ресурс ExPASy (<https://www.expasy.org>). У меню головного пошукового екрана зліва вибираємо опцію «proteomics». Із запропонованого меню вибираємо ресурс «ProtScale». У поле ресурсу вставляємо заздалегідь скопійовану амінокислотну послідовність для лізоциму людини у форматі FASTA та натискаємо «Submit».

ExPASy  
Bioinformatics Resource Portal

ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This program provides 57 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. P06130) or a sequence identifier (ID) (e.g. KPC1\_DROME)

Or you can paste your own sequence in the box below:

```

MKAALTVLGLVLLSTYQGEVFFRCALRTLGLRGLGDFGLSLANNLCAK
KSRQVTEA
TFPNAQRSTGSDTGLQLKSRVWQDQKTFGAVNACLSKSLQLQNTLDAVA
CAEELVYSD
PGLGIRAWTAMWBCQGRDVGIVYGCDF

```

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

- Molecular weight
- Bulkiness
- Number of codon(s)
- Polarity / Zimmerman

4. Отримуємо профіль гідрофобності, з якого видно, що кількість згинів (значення -2) дорівнює 3 (за результатами бази даних UniProt згинів 5).





2. Отримуємо результати, де показані місця дії протеолітичних ферментів на поліпептидний ланцюг. Аналізуємо результати.

### Завдання 3

Використовуючи бази даних SCOP та PROSITE, дослідіть патерни та фолди для каталази людини.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop>) (перехід на базу SCOP можна здійснювати через базу даних ExPASy). На головному пошуковому екрані SCOP вибираємо опцію «top of the hierarchy».

2. Переходимо за посиланням. У пошуковому полі вводимо запит для каталази людини (human catalase) і натискаємо «Search».

Structural Classification of Proteins

Root: scop

Classes:

1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
4. Alpha and beta proteins (a-b) [53931] (376)
5. Multi-domain proteins (alpha and beta) [56572] (66)
6. Membrane and cell surface proteins and peptides [56835] (58)
7. Small proteins [56992] (90)
8. Coiled coil proteins [57942] (7)
9. Low resolution protein structures [58117] (26)
10. Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)

Enter search key:

Generated from scop database 1.7.3 with scopin 1.1.01 on wed Jun 3 10:42:06 2009  
Copyright © 1994-2009 The scop authors / scop@mrc-lmb.cam.ac.uk

3. Отримуємо результати по патернах та фолдах для каталази людини.

Structural Classification of Proteins

Search Results for "human+|-catalase" [scop 1.75]

Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Human (Homo sapiens), cluster 1 [TaxId: 9606]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 1.1 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 1.2 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 2 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 2.2 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 3 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Mouse (Mus musculus), cluster 4 [TaxId: 10090]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Rat (Rattus norvegicus) [TaxId: 10116]  
 Protein: Immunoglobulin light chain kappa variable domain, VL-kappa from Engineered (including hybrid species) [TaxId: 9606]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Human (Homo sapiens), cluster 1 [TaxId: 9606]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 1 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 2 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 2.1 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 2.2 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 3 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 3.1 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 3.2 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Mouse (Mus musculus), cluster 4 [TaxId: 10090]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Rat (Rattus norvegicus) [TaxId: 10116]  
 Protein: Immunoglobulin heavy chain variable domain, VH from Engineered (including hybrid species) [TaxId: 9606]  
 Protein: Immunoglobulin light chain kappa constant domain, CL-kappa from Human (Homo sapiens) [TaxId: 9606]  
 Protein: Immunoglobulin heavy chain gamma constant domain 1, CH1-gamma from Human (Homo sapiens) [TaxId: 9606]  
 Protein: Integrin alpha1-beta1 from Rat (Rattus norvegicus) [TaxId: 10116]

Copyright © 1994-2009 The scop authors / scop@mrc-lmb.cam.ac.uk

4. Здійснюємо вхід на веб-ресурс PROSITE (<https://prosite.expasy.org>). У пошуковому полі PROSITE вставляємо заздалегідь збережену амінокислотну послідовність для катализи людини з бази даних UniProt у форматі FASTA та натискаємо «Scan».

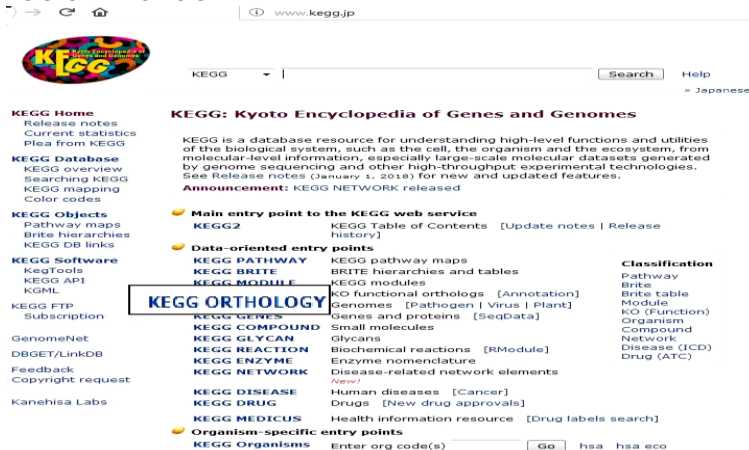
5. Отримуємо результати про специфічні ділянки катализи людини та її домену.

## Завдання 4

Встановить ортологи для цитохрому b<sub>5</sub>.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс KEGG – [www.kegg.jp](http://www.kegg.jp). У меню головного пошукового екран зліва вибираємо опцію «KEGG ORTHOLOGY».



The screenshot shows the KEGG website homepage. The left-hand navigation menu is visible, with the 'KEGG ORTHOLOGY' link highlighted by a red rectangular box. The main content area displays the KEGG logo and a list of navigation links, including 'KEGG Home', 'KEGG Database', 'KEGG Objects', 'KEGG Software', 'KEGG FTP', 'GenomeNet', 'DBGET/LinkDB', 'Feedback', 'Copyright request', and 'Kanehisa Labs'. The main content area also features a search bar and a list of KEGG categories, with 'KEGG ORTHOLOGY' highlighted in a red box.

2. Переходимо на сторінку KEGG ORTHOLOGY та в пошуковому полі вводимо запит «cytochrome b<sub>5</sub>».



The screenshot shows the KEGG KO (KEGG ORTHOLOGY) Database search page. The search bar contains the query 'cytochrome b5'. The page header includes the KEGG logo and the text 'KO (KEGG ORTHOLOGY) Database'. Below the search bar, there is a 'Go' button. The page also features a navigation menu with options like 'Menu', 'PATHWAY', 'BRITE', 'MODULE', 'KO', 'Annotation', 'ENZYME', 'RModule', and 'BlastKOALA'.

### KO Database of Molecular Functions

In KEGG, molecular-level functions are stored in the **KO (KEGG Orthology)** database and associated with ortholog groups in order to enable extension of experimental evidence in a specific organism to other organisms. Genome annotation in KEGG is ortholog annotation, assigning KO identifiers (K numbers) to individual genes in the GENES database. No updates are made to original data, such as gene names and descriptions given by RefSeq or GenBank, even if they are inconsistent with the KO assignment.

Major efforts have been initiated to associate each KO entry with experimental evidence of functionally characterized sequence data, now shown in the SEQUENCE subfield of the REFERENCE field. Furthermore, the genome-based collection of KEGG GENES has been expanded to allow individual protein data to be included in the addendum category. Eventually the KO database will cover all knowledge on functionally characterized protein sequences (see also KEGG Enzyme).

### KEGG Mapping by the KO System

In general KO grouping of functional orthologs is defined in the context of KEGG molecular networks (KEGG pathway maps, BRITE hierarchies and KEGG modules), which are in fact represented as networks of nodes identified by K numbers. The relationships between KOs and corresponding molecular networks are represented in the following KO system.

### 3. Отримуємо запит про всі ортологи для цитохрому b5.

The screenshot shows a web browser window with the URL [www.kegg.jp/dbget-bin/www\\_bfind\\_sub?mode=bfind&max\\_hit=1000&dbkey=orthology&keywords=cytochr](http://www.kegg.jp/dbget-bin/www_bfind_sub?mode=bfind&max_hit=1000&dbkey=orthology&keywords=cytochr). The search results are displayed in a table with the following entries:

Accession	Description
K00236	SDHC, SDH3, succinate dehydrogenase (ubiquinone) cytochrome b560 subunit
K00326	E1.6.2.2, cytochrome-b5 reductase [EC:1.6.2.2]
K02707	psbE, photosystem II cytochrome b559 subunit alpha
K02708	psbF, photosystem II cytochrome b559 subunit beta
K08360	CYB561, cytochrome b-561 [EC:1.16.5.1]
K08371	CYB561D2, cytochrome b-561 domain containing protein 2
K12262	cytB, cytochrome b561
K15536	cytC, soluble cytochrome b562

DBGET integrated database retrieval system

### Завдання 5

Використовуючи програму RasMol, дослідіть структуру гемоглобіну людини. У молекулі гемоглобіну людини виділіть: а) білкову частину; б) гем; в) атом заліза; г) серини білкової частини; д) гістидини, розміщені біля гему; е) одну із субодиниць; є) водневі зв'язки.

### Рекомендація

1. На персональний комп'ютер завантажуюмо програму RasMol, доступну на сайті <http://www.openrasmol.org>. На головному екрані програми натискаємо «RasMol 2.7.5. Windows Installer» і завантажуюмо програму на комп'ютер.

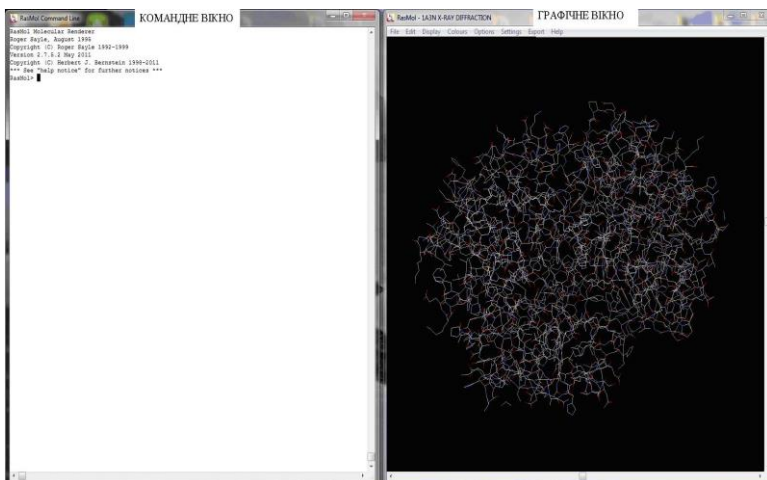
The screenshot shows the homepage of the RasMol and OpenRasMol website. The page features a navigation menu with links to 'Copying and Distribution', 'Contents', 'Software Distributions', 'Latest Windows Installer', and 'External Packages'. The main heading is 'Home Page for RasMol and OpenRasMol', with the subtitle 'Molecular Graphics Visualisation Tool'. Below this, there are two columns of links. The left column includes 'RasMol Latest Windows Installer', 'RasMol Latest Source Tarball', 'RasMol Latest Manual', and 'Donate to Support RasMol'. The right column includes 'RasMol 2.7.5 Windows Installer', 'RasMol 2.7.5 Source Tarball', 'RasMol 2.7.5 Manual', and 'Donate to Support RasMol'. In the center, there is a 3D molecular model of a protein structure.

2. У базі даних PDB (<https://www.rcsb.org>) знаходимо 3D-структуру гемоглобіну людини (human hemoglobin) з ID 1A3N. Справа вверху натискаємо «Download Files», вибираємо формат «PDB Format (gz)» та зберігаємо білок на комп'ютер.

The screenshot shows the RCSB PDB website interface. At the top, there is a navigation bar with options like 'Deposit', 'Search', 'Visualize', 'Analyze', 'Download', 'Learn', and 'More'. Below this is the PDB logo and search bar. The main content area is for entry 1A3N, 'DEOXY HUMAN HEMOGLOBIN'. A 'Download Files' dropdown menu is open, showing options: 'FASTA Sequence', 'PDB Format', 'PDBmmCIF Format', 'PDBmmCIF Format (gz)', and 'PDBML/XML Format (gz)'. An arrow points to 'PDB Format (gz)'. To the right, there is a 'wwPDB Validation' section with a table of metrics.

Metric	Biological Assembly 1
Rfree	Structure Factors (CIF)
Clashscore	Structure Factors (CIF - gz)
Ramachandran outliers	
Sidechain outliers	

3. Відкриваємо програму RasMol. Робота з програмою відбуватиметься у двох вікнах: графічному та командному. У графічне вікно вставляємо 3D-структуру гемоглобіну людини (перетягнувши її з попередньо збереженого файлу).





4. У командному вікні набираються дії, які відповідають відповідним командам, кожна з яких набирається англійською з клавіатури при активному командному вікні і завершується натисканням клавіші «Enter». Позначення деяких команд подано в таблиці.

Команди	Значення
select «безліч»	виділяє безліч
restrict «безліч»	виділяє безліч і стирає з графічного вікна все інше
wireframe 50	додає до зображення в графічному вікні дротяну модель виділеної безлічі з товщиною лінії 50
wireframe off	стирає з графічного вікна дротяну модель виділеної безлічі
backbone 70	додає до зображення в графічному вікні остовну модель виділеної безлічі з товщиною лінії 70
cpk 200	додає до зображення в графічному вікні кулькову модель виділеної безлічі з діаметром кульок 200
color «колір»	зафарбовує безліч у вказаний колір (якщо атоми не були зображені в жодній моделі, то колір не буде видно, доки ви їх не виділите).

5. Виконаємо завдання для візуалізації гемоглобіну людини:

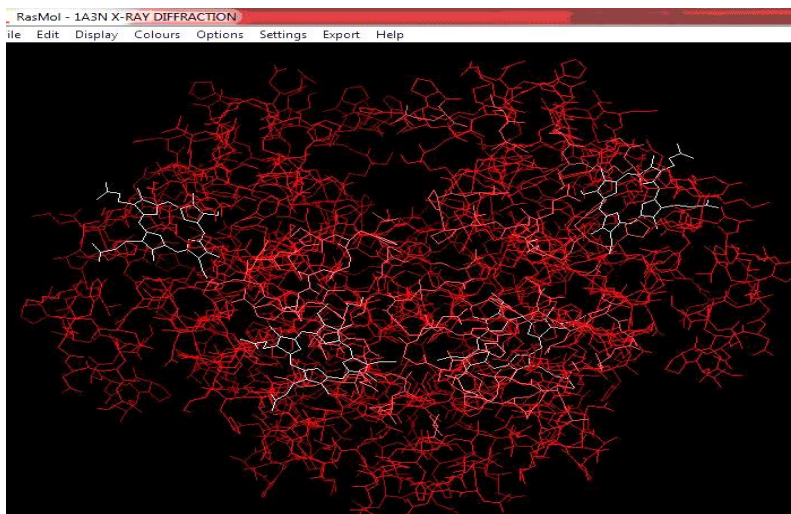
а) для виділення білкової частини гемоглобіну та позначення її червоним кольором у командному вікні задаємо команди «select protein» → «color red».

```

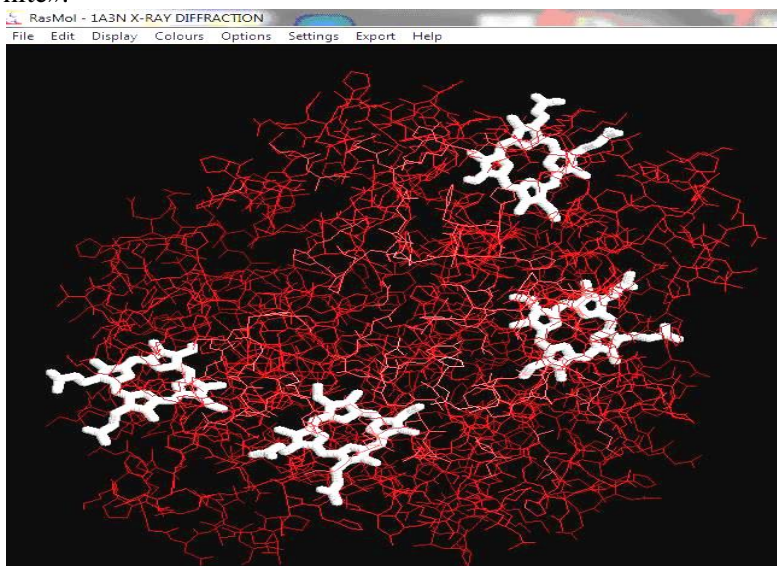
RasMol Command Line
RasMol Molecular Renderer
Roger Sayle, August 1995
Copyright (C) Roger Sayle 1992-1999
Version 2.7.5.2 May 2011
Copyright (C) Herbert J. Bernstein 1998-2011
*** See "help notice" for further notices ***
RasMol>
RasMol> select protein
4370 atoms selected!
RasMol> color red
  
```



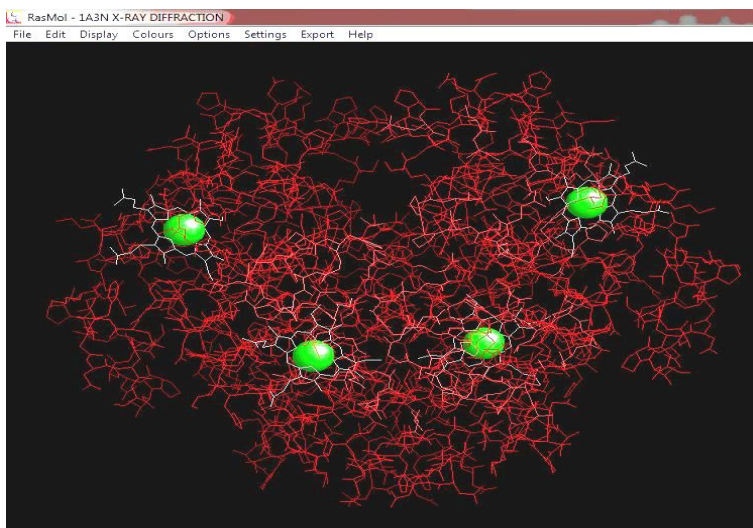
Білкова частина зафарбовується у червоний колір.



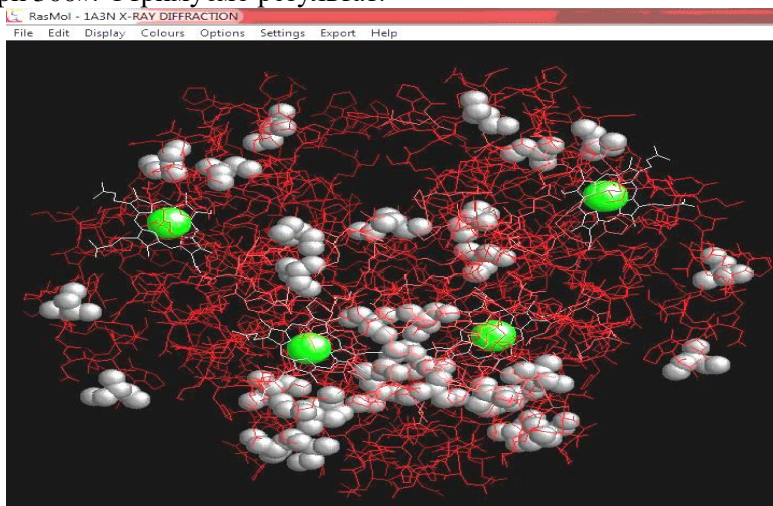
б) виділяємо гем та зафарбовуємо його у білий колір, для цього задаємо команди «select hem» → «wireframe 100» → «color white».



в) атом заліза виділяємо у формі кульки діаметром 500. У командному вікні задаємо: «select iron» → «color green» → «cprk 500». Отримуємо результат.



г) виділяємо серини білкової частини у формі кульок діаметром 50, задавши команди: «select Ser» → «color grey» → «cprk 300». Отримуємо результат.

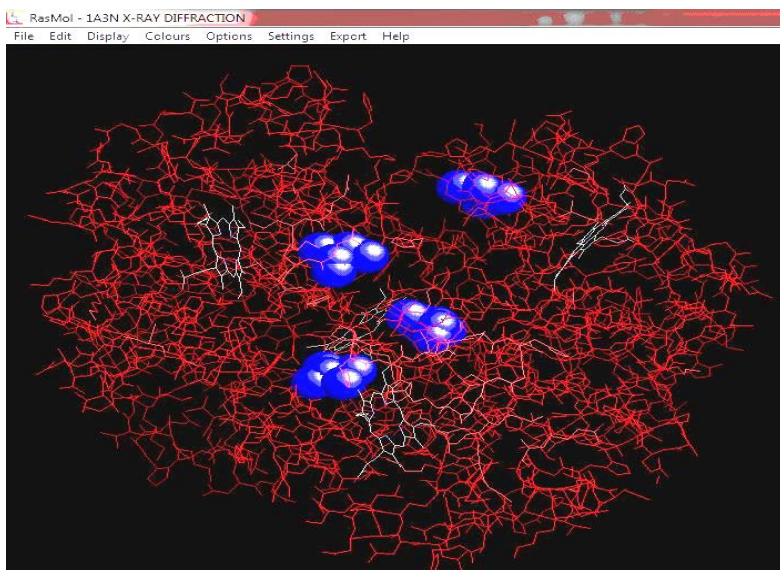


д) щоб, виділити гістидини, розміщені біля гему, в базі даних Uniprot (<http://www.uniprot.org>) з'ясуємо їхнє положення в поліпептидному ланцюзі. Для цього заходимо на результат пошуку гемоглобіну людини з ідентифікатором P68871 і встановлюємо, що гістидини, пов'язані із залізом гему, знаходяться в положеннях 64 та 93.

The screenshot shows the Uniprot entry for HBB (Homo sapiens). The 'Function' section is expanded, showing a table of binding sites. Two specific sites are highlighted with boxes:

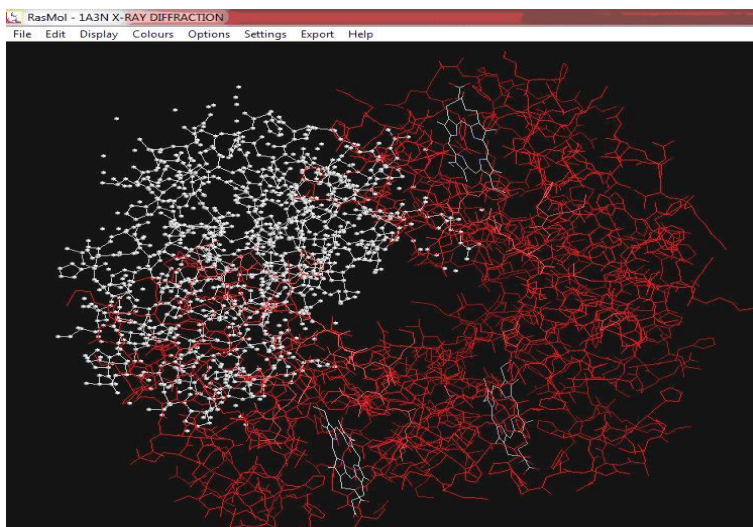
Feature key	Position(s)	Description
Binding site <sup>1</sup>	2	2,3-bisphosphoglycerate; via amino nitrogen
Binding site <sup>1</sup>	64	Iron (heme distal ligand)
Metal binding <sup>1</sup>		
Binding site <sup>1</sup>	93	Iron (heme proximal ligand)
Metal binding <sup>1</sup>		
Binding site <sup>1</sup>		

Для їх виділення задаємо команди «select his64, 93» → «color blue» → «crk 500». Отримуємо результат.

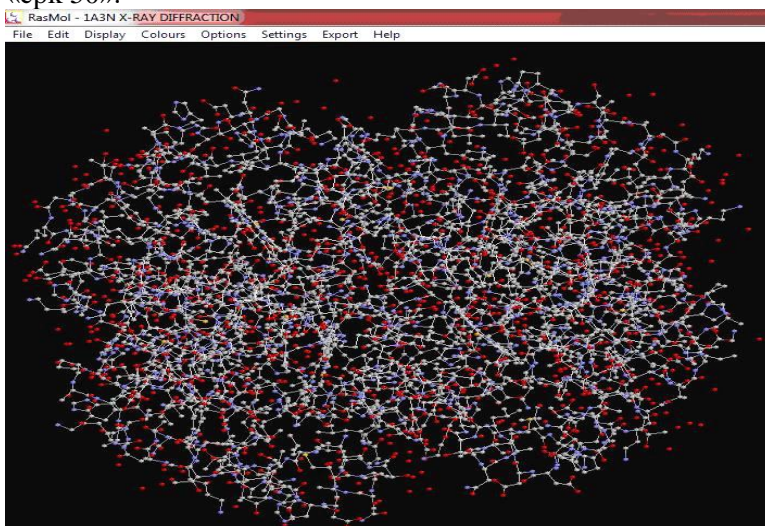




е) виділяємо всі атоми одного з ланцюгів гемоглобіну у формі кульок діаметром 40. Задаємо команди «select \*A» → «srk 50» → «color white». Отримуємо результат.



є) водневі зв'язки виділяємо командою: «color hbonds green» → «srk 50».



---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

**1. Який елемент вторинної структури білка допомагає передбачити побудова профілю гідрофобності?**

- а) спіраль;
- б) складчастий лист;
- в) згин;
- г) дисульфідні зв'язки.

**2. Який із біоінформаційних підходів дає змогу отримати максимально реальну тривимірну структуру білкової молекули?**

- а) передбачення вторинної структури та фолдів;
- б) накладання структур та структурні вирівнювання на основі шаблону;
- в) розпізнавання фолду на основі амінокислотної послідовності;
- г) моделювання de novo та моделювання «з перших принципів».

**3. Розбіжності гемоглобіну та міоглобіну, які виконують в організмі різні функції, – приклад .....**

- а) функціональної конвергенції білків;
- б) функціональної дивергенції білків.

**4. Карти Рамачандрана сприяють виявленню .....**

- а) четвертинної структури;
- б) третинної структури;
- в)  $\alpha$ -спіралі;
- г)  $\beta$ -листів.

**5. Які методи допомагають встановити елементи вторинної структури?**

- а) побудова профілю гідрофобності;
- б) розщеплення поліпептиду протеазами;
- в) парне вирівнювання послідовностей;
- г) побудова карти Рамачандрана.

**6. Карти Рамачандрана використовуються для демонстрації наявності у білку .....**

- а) дисульфідних зв'язків;
- б) водневих зв'язків;
- в) вторинних структур;
- г) третинної структури.

**7. Заповніть пропуски: *Процес, що призводить до формування комплексу схожих ознак у представників неспоріднених груп білків називається .....***

---

---

**8. Який метод моделювання тривимірної структури білків використовує систему оцінювання сумісності послідовності-мішені з кожним із фолдів у бібліотеці баз даних?**

- а) моделювання білків de novo;
- б) моделювання «з перших принципів»;
- в) метод протягування.

**9. Спосіб просторового розміщення поліпептидного ланцюга у тривимірній структурі білка називається .....**

- а) фармакофором;
- б) паралогом;
- в) патерном фолдингу;
- г) лідом.

**10. Чи правильне твердження: У базі даних PDB не можуть траплятися повністю помилкові моделі, оскільки при внесенні інформації в базу даних вона перевіряється кураторами?**

- а) так; б) ні.

**11. Стандарти значення яких кутів ураховують при побудові карти Рамачандрана?**

- а) « $\omega$ »;
- б) « $\phi$ »;
- в) « $\psi$ »;
- г) « $\chi$ ».

**12. Доповніть речення: Методи моделювання структури білків об'єднуються у два типи – а) .....; б) .....**

**13. Чи правильне твердження: На практиці карти Рамачандрана використовуються для демонстрації наявності у білку вторинних і надвторинних структур?**

- а) так; б) ні.

**14. Виберіть основні недоліки, притаманні методу моделювання ab initio:**

- а) ефективні лише для поліпептидів, які містять понад 100 амінокислотних залишків;
- б) ефективні лише для поліпептидів, які містять менш як 100 амінокислотних залишків;
- в) надзвичайно вимогливі до обчислювальних ресурсів.

---

---

**15. У базі даних PDB знайдено шаблон-гомолог досліджуваного білка. Який метод моделювання доречно виконувати у цьому разі?**

- а) гомологічне моделювання;
- б) моделювання методом «протягування»;
- в) моделювання de novo;
- г) моделювання «з перших принципів».

**16. Заповніть пропуски: *Процес, за допомогою якого білки, після дуплікації генів, набувають функцій, відмінних від предкового білка, називається .....***

**17. Найефективнішим методом моделювання тривимірної структури білків є .....**

- а) моделювання білків de novo;
- б) моделювання «з перших принципів»;
- в) гомологічне моделювання.

**18. Доповніть речення: *Особливий згорнутий стан або структура білка, яка має низьку енергію, тобто є термодинамічно стійкою при фізіологічних умовах, називається .....***

**19. Які методи найширше використовують у передбаченні вторинної структури білків?**

- а) метод Чоу-Фасмена;
- б) моделювання білків de novo;
- в) гомологічне моделювання;
- г) методи моделювання нейронних мереж;
- д) моделювання методом «протягування».

**20. Розставте методи передбачення третинної структури білка від найбільш до найменш ефективних:**

- а) моделювання білків de novo;
- б) моделювання «з перших принципів»;
- в) гомологічне моделювання;
- г) рентгеноструктурний аналіз.

**21. Заповніть пропуски: ..... – це один із ряду конформерів, який є результатом обмеженого обороту навколо єдиного хімічного зв'язку.**

**22. Розставте етапи гомологічного моделювання в хронологічному порядку .....**

- а) оптимізація моделі;

- 
- 
- б) моделювання петель;
  - в) створення поліпептидного каркасу;
  - г) ідентифікація шаблонів та первинне вирівнювання послідовностей;
  - д) корекція вирівнювання послідовностей;
  - е) валідація моделі;
  - ж) модель цільового білка.

**23. У яких базах даних можна знайти спеціальні бібліотеки фолдів білків, які використовують для їхнього моделювання?**

- а) CATH;
- б) EMBL;
- в) SCOP;
- г) GOBASE.

**24. Якщо в базі даних PDB не знайдено шаблон-гомолог досліджуваного білка, який метод моделювання потрібно застосувати?**

- а) гомологічне моделювання;
- б) моделювання методом «протягування»;
- в) моделювання de novo;
- г) моделювання «з перших принципів».

**25. Програма Ramachandran Plot Explorer допомагає .....**

- а) виконати моделювання тривимірної структури білка;
- б) аналізувати конформаційні зміни;
- в) аналізувати вторинну структуру та загальну структуру петель.

**26. Виберіть програми, призначені для візуалізації молекул і отримання зображень просторових структур біологічних макромолекул .....**

- а) BLAST;
- б) DALI;
- в) Clustal;
- г) T-Coffee;
- д) RasMol;
- е) Jmol.

**27. Чи дає змогу карта Рамачандрана однозначно встановити третинну структуру невідомого білка?**

- а) так; б) ні.



---

---

**28. Ферменти гексокіназа, рибокіназа і галактокіназа проявляють подібні ферментативні функції, однак мають різну тривимірну структуру. Цей приклад відображає:**

- а) функціональну конвергенцію білків;
- б) функціональну дивергенцію білків.

**29. Заповніть пропуски: *Фізико-хімічний процес, в результаті якого білки в своєму природному середовищі (розчині, цитоплазмі або мембрані) набувають характерного лише для них просторового укладання і функцій, називається .....***

**30. У якій програмі передбачення структури білка реалізований ієрархічний алгоритм – згортання починається із коротких фрагментів, поступово об'єднуючи їх у довші?**

- а) Ramachandran Plot Explorer;
- б) Modeller;
- в) LINUS.

---

---

## ТЕМА 5. БІОІНФОРМАТИКА У ПОШУКУ ТА РОЗРОБЦІ ЛІКІВ

**Мета:** ознайомитися з етапами пошуку та розробки ліків. Навчитися систематизувати різні підходи до розробки ліків із використанням комп'ютерного моделювання.

### ТЕОРЕТИЧНА ЧАСТИНА

#### 1. Основні етапи пошуку та розробки ліків

Протягом тривалого часу основним джерелом сполук для експериментального тестування на фізіологічну активність були живі організми (рослини, тварини, мікроорганізми) і класичний хімічний синтез. При експериментальному тестуванні велика кількість сполук відкидалася як неперспективні прототипи ліків через: низьку цільову активність (або повну її нестачу), токсичність, канцерогенність, складність синтезу. Тільки одна зі 100 тисяч сполук могла стати лікарським препаратом із виходом на фармацевтичний ринок.

Шлях створення нового лікарського препарату має сім основних етапів:

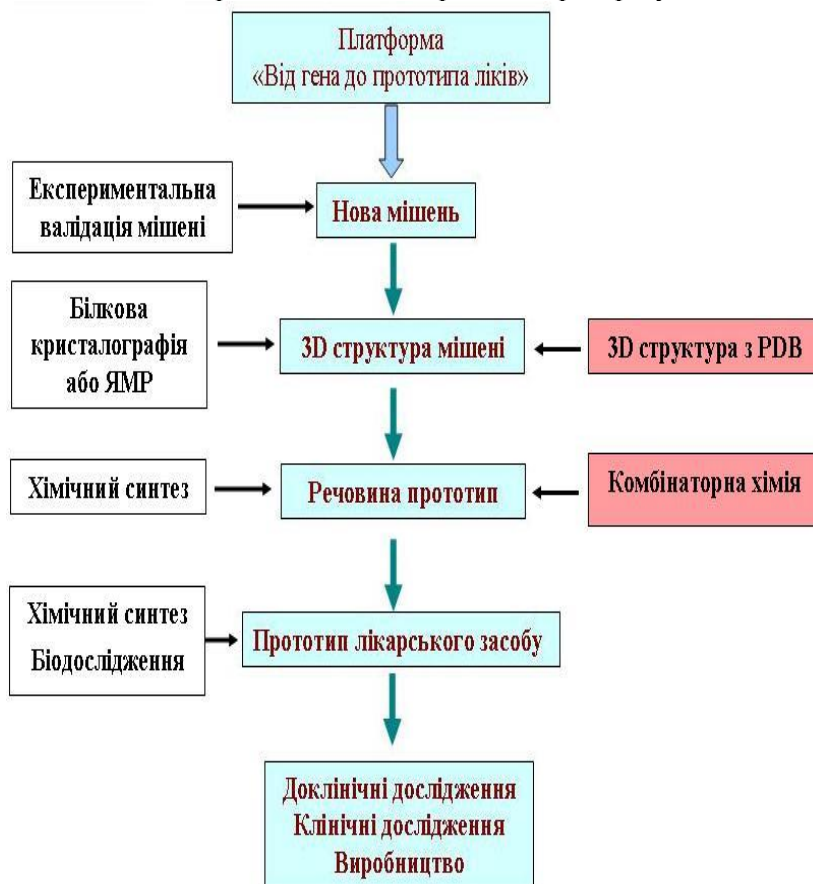
1. Вибір хвороби, для лікування якої створюється лікарський засіб.
2. Вибір молекулярної мішені для лікарського засобу.
3. Знаходження базової структури для нового лікарського засобу (речовини прототипу).
4. Оптимізація базової структури та створення прототипу лікарського засобу.
5. Доклінічні дослідження.
6. Клінічні дослідження.
7. Виробництво препарату.

Загальні витрати часу і грошей могли сягати від 12 – 15 років і понад 800 млн доларів. Скоротити час і фінансові затрати на останніх етапах (клінічні дослідження і створення препарату) фактично неможливо з причин суворих державних стандартів і законів. Тому основні зусилля розробників ліків, спрямованих на підвищення ефективності процесу створення нових препаратів, стосуються більш ранніх стадій.

За останнє десятиріччя сучасні комп'ютерні технології, біоінформатика та нові експериментальні методи у галузі

медичної хімії об'єднані в загальний підхід, названий раціональним використанням ліків, який забезпечує прискорення й оптимізацію процесу знаходження нових біологічно активних сполук – базових структур нових ліків. Крім того, засоби біоінформатики допомагають передбачати багато нових мішеней (білків, генів) для дії ліків.

Тому інтеграція зазначених підходів зумовила створення єдиної платформи «Від гена – до прототипа ліків», яка охоплює 4 із 7 етапів створення нового лікарського препарату.



*Рис. 37. Етапи створення нового лікарського препарату*

---

Як видно з рис. 37, деякі блоки цієї платформи містять комп'ютерне моделювання. Виявлення й аналіз мішені для лікарського препарату можливі за допомогою експериментальних методів (ЯМР, кристалографія) або ж комп'ютерного моделювання. Завдяки засобам біоінформатики виконують віртуальний скринінг хімічних речовин та передбачають властивості речовин, що допомагає знизити затрати часу та фінансування.

Методи комп'ютерного конструювання ліків дають змогу скоротити на два порядки кількість сполук, які необхідно синтезувати та перевірити на наявність цільової біологічної активності. Однак ці підходи не можуть повністю замінити реальні експерименти.

Мета комп'ютерних методів – генерація високоймовірних гіпотез про нові мішені дії ліків та ліганди, які взаємодіють із мішенями (основа для майбутніх ліків). У майбутньому ліганди повинні бути перевірені в прямих експериментах.

Для розробки лікарських препаратів спочатку визначають хворобу-мішень, яка має свої специфічні симптоми, генетичні передумови, епідеміологію, її взаємозв'язок з іншими захворюваннями людини і тварин та усі відомі на сьогодні засоби лікування. Із хвороби-мішені вибирають мішень. Більшість фармакологічних мішеней – білки, яких на сьогодні налічується близько 1000, однак вважається, що загальна їхня кількість становить 10000. Найбільша частина відомих фармакологічних мішеней (близько 50 % від загальної кількості) – рецептори клітинних мембран. Приблизно 30 % – ферменти. Важливу групу становлять іонні канали і білки-переносники нейромедіаторів і гормонів (дофамін, серотонін та ін.).

Водночас паралельно необхідно розробляти засоби аналізу для визначення дієвості застосованого підходу.

Основна мета ранніх стадій розробки ліків – встановлення однієї чи більше лідерних сполук – *лідів*. Це може бути будь-яка субстанція, яка проявляє біологічну активність, яку шукають. Класичні приклади лідів – це:

- природні речовини: ванкоміцин (антибіотик проти бактеріальної інфекції), стауроспорин (лід для розробки селективних інгібіторів кіназ проти раку), рапаміцин (для

- 
- 
- імуносупресії), таксол (протипухлинний агент);
  - синтетичні молекули: прозак (від депресії), ліпітор (від гіперхолестеролемії), глівек (від хронічної лейкоїмії).

*До шляхів пошуку лідів належать:*

1. Інтуїція. Класичний приклад – пеніцилін.
2. Скринінг природних джерел, наприклад дігіталіс, виділений із наперстянки, який застосовувався для лікування серцевих захворювань. Близько половини ліків, які використовуються в наш час, створені на основі природних сполук.
3. Аналіз уже відомої інформації про субстрати та механізми їхньої дії, а також вибір активних компонентів із цих сполук.
4. Використання ліків, ефективних проти споріднених захворювань.
5. Скринінг сполук засобами комбінаторної хімії.
6. Винайдені випадково препарати, з побічних ефектів наявних ліків. Наприклад, міноксидил спершу розроблений як антигіпертензивний засіб, викликаючи згодом посилення росту волосся. Віагра також спочатку розроблялася як засіб для лікування серцево-судинних захворювань. Тамоксифен, який нині широко використовується проти раку молочної залози, спочатку передбачався як контрацептив (він дійсно гарний контрацептив для шурів, але викликає овуляцію у жінок).
7. Експериментальний скринінг.
8. Комп'ютерний скринінг та *ab initio*-комп'ютерний дизайн.

Використання досліджень для відбору та винайдення ліків великою мірою залежить від наявності різноманітних хімічних бібліотек, які створюються методами комбінаторної хімії. Такі бібліотеки спочатку підвищують можливість пошуку молекул, які будуть взаємодіяти з певною білковою мішенню.

## **2. Основні стратегії комп'ютерного моделювання ліків**

Комп'ютерний та інформаційний пошук робить свій внесок у кроки на шляху до розробки ліків.

До основних етапів біоінформаційної розробки ліків належать: ідентифікація мішені; моделювання, аналіз та добір лід-лігандів; вибір і скринінг бібліотек.

Для розробки ліків проти певного захворювання важливо

визначити білок-мішень, пов'язаний із цим захворюванням, при впливі на функціонування чи експресію якого вдається досягнути бажаного лікувального ефекту.

Хімічна сполука (ліганд), яка при взаємодії з рецептором-мішенню змінює його стан, сприяючи біологічній відповіді, називається агоністом. *Звичайні агоністи* підвищують відповідь рецептора, *зворотні* – знижують її. Антагоністи блокують дію агоністів. Агоністи й антагоністи – модулятори фармакологічних мішеней, які поділяються на позитивні та негативні.

**Таблиця 3**

***Класифікація модуляторів фармакологічних мішеней***

<i>Біомолекули</i>	<i>Позитивні модулятори</i>	<i>Негативні модулятори</i>
ферменти	активатори	інгібітори
рецептори	агоністи	антагоністи
ьонні канали	деблокатори	блокатори

Залежності від того, в якому сигнальному шляху бере участь конкретний рецептор, агоністи зумовлюють різну відповідь. Для розуміння зв'язування антагоністів і зворотних агоністів із рецепторами застосовують комп'ютерне моделювання структур білка-рецептора на основі вже відомих структур білків.

Стратегія пошуку біологічно активних молекул значною мірою залежить від того, відомі чи ні тривимірні структури молекули-біоліганда та рецептора-мішені. При цьому коштовнішим є знання структури рецептора, що дає змогу здійснювати пряме моделювання. Таблиця 4 відображає можливі поєднання знань про структуру ліганда та рецептора й основні підходи до розробки ліків у кожному з випадків.

**Таблиця 4**

***Основні стратегії, які використовують у комп'ютерному моделюванні ліків***

<b>Структура рецептора</b>	<b>Структура ліганда</b>	
	<i>Відома</i>	<i>Невідома</i>
<i>відома</i> (прямий дизайн)	докінг	дизайн de novo
<i>невідома</i> (непрямий дизайн)	аналоговий дизайн (QSAR)	фармакофорний скринінг

---

Стисло охарактеризуємо названі методи.

**Пряме моделювання** – один із найефективніших підходів під час пошуку лікарських речовин. Це зрозуміло, оскільки в його рамках відтворюється структура лігандорецепторного комплексу з оцінкою конформацій і взаємної спорідненості.

За допомогою *алгоритму докінгу* можна передбачити взаємодію молекул, визначити просторову структуру їхніх комплексів та афінність взаємодії. Алгоритми докінгу низькомолекулярних сполук до молекул рецептора – невід’ємне знаряддя для раціонального дизайну ліків.

Завдяки докінгу виконують молекулярно-графічний аналіз комплексу лігандорецептора (можливість розміщення потенційного ліганда в порожнині активного центра макромолекули) й оцінку енергії зв’язування та спорідненості для такого комплексу.

Докінг може застосовуватися на різних стадіях процесу розроблення ліків. Найчастіше його використовують для віртуального скринінгу бібліотек низькомолекулярних органічних сполук і фрагментів. Для прогнозування токсичності потенційні ліки можна тестувати *in silico* на метаболічно-активних протеїнах. Отже, комп’ютерне фільтрування допомагає зберегти не лише багато тестів *in vitro*, але й дорогі експерименти *in vivo*.

Докінг-процедури оцінюють комплемент відомих структур до активного центра. Для пошуку зручно використовувати наявні бази даних відомих сполук, наприклад ChEMBL (<https://www.ebi.ac.uk/chembl>). Розроблено безліч програмних пакетів для практичної реалізації цього підходу. Найпопулярніші програми молекулярного докінгу доступні через Інтернет: FlexX (<https://www.biosolveit.de/FlexX>); GOLD (<https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold>); AutoDock (<http://autodock.scripps.edu>); LIGPLOT ([https://www.researchgate.net/post/How to use LIGPLOT software build a diagram like this2](https://www.researchgate.net/post/How_to_use_LIGPLOT_software_build_a_diagram_like_this2)); Hex (<http://hex.loria.fr>); FRED (<https://docs.eyesopen.com/oedocking/fred.html>).

Основні методологічні труднощі докінгу пов’язані з обліком конформацій ліганда, гнучкості рецептора і побудовою функції.

У *методах de novo* структура цільових молекул відтворюється поступовим конструюванням на основі невеликих

---

---

фрагментів, вміщених на активний сайт рецептора. При цьому мінімізують енергію відштовхування груп (стеричний чинник) і максималізують енергію зв'язування.

Використані комп'ютерні процедури будують гіпотетичні структури, які повинні мати високу спорідненість із рецептором. Недоліки цієї методології – відносно невисока надійність оцінки спорідненості лігандів до мішені і дуже загальні правила побудови молекул (значна частина структур, які генеруються, не може бути синтезована на практиці).

**Непрямі підходи** базуються переважно на побудові залежностей «структура-активність» (Quantitative Structure-Activity Relationship (QSAR)) – аналоговий дизайн та фармакофорний аналіз.

**Аналоговий дизайн.** Багато компонентів із фармакологічною активністю здатні проявляти споріднену активність, відрізняючись потенціалом і специфічністю. Починаючи з ліда, хіміки повинні провести скринінг великої кількості споріднених сполук, щоб оптимізувати необхідні фармакологічні властивості. Саме кількісне співвідношення структура-активність забезпечується методами для передбачення фармакологічної активності набору компонентів на основі взаємозв'язку між молекулярними властивостями та фармакологічною активністю.

При цьому аналізі передбачається, що основний вклад у лігандорецепторні взаємодії вносять деякі функціональні групи, тому їх можна вважати відповідальними за взаємозв'язок структури й активності.

Аналіз охоплює два етапи: виявлення фармакофорних груп і просторове поєднання активних конформацій низки молекул або їхніх силових полів для кращого розуміння ключових структурних особливостей. Нині розроблена значна кількість програм для практичної реалізації даного підходу.

QSAR застосовують у моделюванні біологічної активності вже понад 40 років.

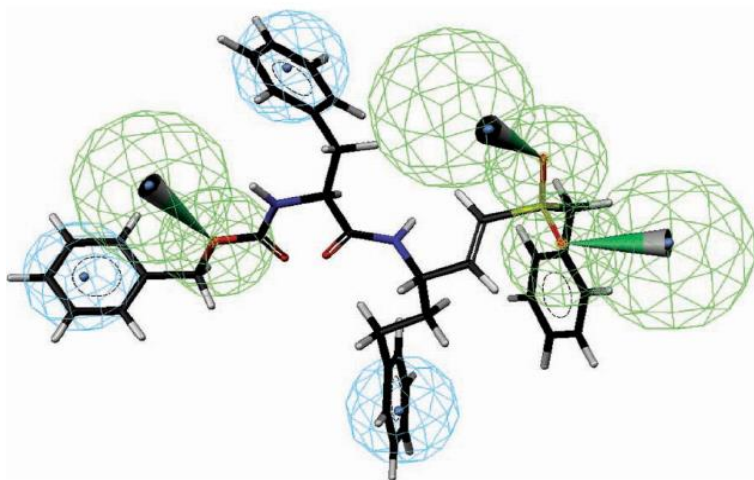
**Фармакофорний скринінг** дає змогу значно скоротити час пошуку необхідних лідів. Перш ніж виконувати експерименти в лабораторних умовах, доцільно зібрати якнайбільше інформації про потенційні взаємодії препарату з мішенню.

Фармакофор – це набір структурних ознак у молекулі, які



---

розпізнаються біологічними рецепторами і відповідальні за біологічну активність молекули. Функціональні групи фармакофора на рис. 38 зображені у формі сфер різного радіуса.



*Рис. 38. Модель фармакофора з молекулою ліда*

Комп'ютерні програми використовують різні кольори для ілюстрації сфер, які відповідають гідрофобним і гідрофільним ділянкам, або позитивно і негативно зарядженим функціональним групам. Вузько спрямовані водневі зв'язки зображені конусами. Відповідна молекула ліда вписана у фармакофорну модель.

Фармакофорний скринінг можна реалізувати за допомогою автоматичного вибіркового пошуку в хімічних базах даних.

**Передбачення ліда.** Методи добору лігандів як лідів для розробки ліків можуть бути об'єднані в індуктивні та дедуктивні підходи.

Індуктивні методи залежать від кореляцій між відомими афінностями деяких тестових наборів компонентів і молекулярними особливостями, які характеризують цілі бібліотеки потенційних лігандів. Ці ознаки поєднують структурні особливості, як-от: розмір, геометрія, розподіл заряду та специфічні функціональні групи, зокрема донори і акцептори протонів, а

---

---

також загальні «ознаки ліків», наприклад розчинність у водних та неполярних розчинниках, легкий шлях введення, необхідний розподіл у тканинах організму та метаболічний шлях виведення. Аналіз такої інформації здійснює хемоінформатика – застосування біоінформатики у розробці ліків.

Дедуктивні методи використовують, якщо відомі сайти зв'язування на білковій молекулі.

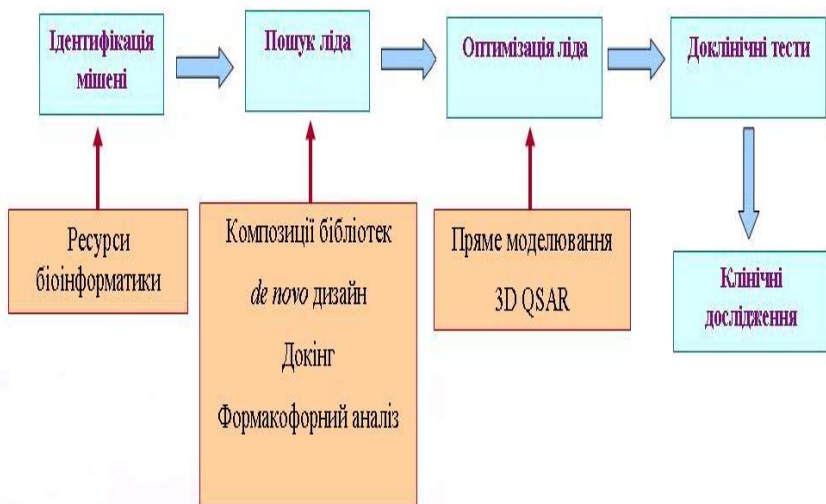
Комп'ютерне моделювання біологічної активності ще не гарантує нового лікарського засобу, проте помітно деталізує та прискорює цей процес, зокрема на стадії оптимізації сполуки-ліда.

### **3. Модернізація ключових етапів процесу розробки лікарських засобів**

Головне завдання в розробці ліків – встановлення компонента, який буде міцно та специфічно зв'язуватися з білком-мішенню. Щільне зв'язування необхідне для ефективності при низьких концентраціях, а специфічність необхідна для мінімізації побічних ефектів.

Якщо структура мішені відома з експерименту, то можна застосовувати молекулярне моделювання для прямої розробки ліганда. Якщо структура мішені невідома – вигляд сайта зв'язування повинен бути змодельований із непрямих доказів.

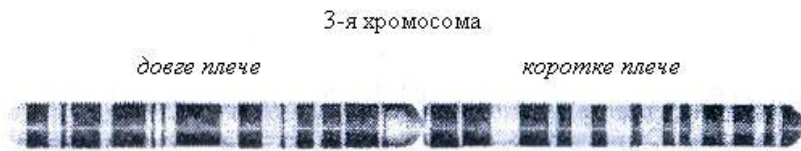
Як зазначалось раніше, основний внесок у оптимізацію процесу розробки лікарського препарату можна зробити на стадії доклінічних випробувань. Традиційно на доклінічній фазі створення лікарського препарату виділяють такі основні етапи: ідентифікацію і валідацію мішені, пошук сполуки-лідера, оптимізацію базової структури та доклінічну оцінку фармакологічних властивостей. Кожна з цих стадій дуже збагатилася як деякими експериментальними процедурами, так і методами комп'ютерного моделювання.



**Рис. 39.** Доповнення різних етапів доклінічної фази розробки лікарських засобів методами комп'ютерного моделювання

### Приклади застосування біоінформатики у розробці ліків

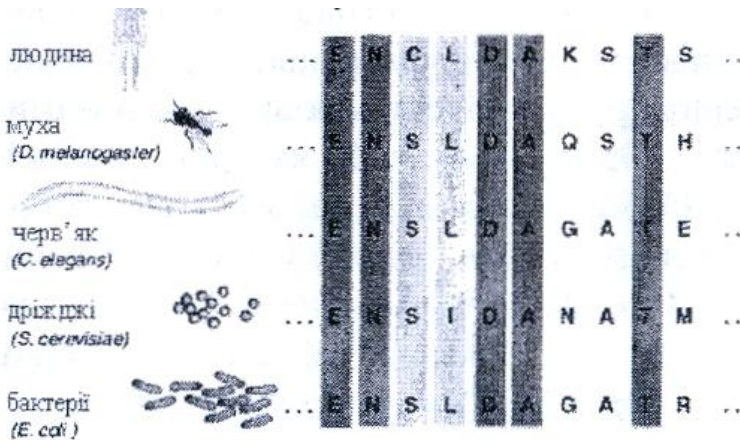
Одним із найперших випадків використання біоінформатики з медичною метою було раціональне моделювання ліків. Один із таких підходів використання продукту гена MLH1 як мішені для розробки ліків. MLH1 – ген людини, який кодує mismatch repair protein (mrr) та розміщений в короткому плечі 3-ї хромосоми.



Методом порівняльного аналізу на основі подібності гена mrr мишей встановлено залучення даного гена до розвитку раку товстої кишки. На основі нуклеотидної послідовності визначена можлива амінокислотна послідовність.

У подальшому техніка встановлення послідовностей

застосовується для пошуку гомологів у модельних організмах і, враховуючи подібність, змодельовують структуру людського білка на основі експериментально охарактеризованих структур.



Використовуючи певні алгоритми, можна сконструювати молекули, які можуть зв'язуватися з модельною структурою, відкриваючи шлях для біохімічного дослідження їхньої біологічної активності щодо справжнього білка.

Постійне вдосконалення процесу розробки лікарських препаратів зумовлене його яскраво вираженою міждисциплінарністю та можливістю залучення новітніх досягнень у суміжних галузях: геноміці, протеоміці, біохімії, молекулярній біології, медицині, фармакології, комп'ютерному моделюванні.

Бурхливий розвиток медичної хімії створює передумови для переходу від методу проб і помилок до дійсно раціонального дизайну ліків, коли емпіричний синтез із поступовою перевіркою активності поступається місцем спрямованому створенню речовин із бажаними фізико-хімічними властивостями та біологічною дією.

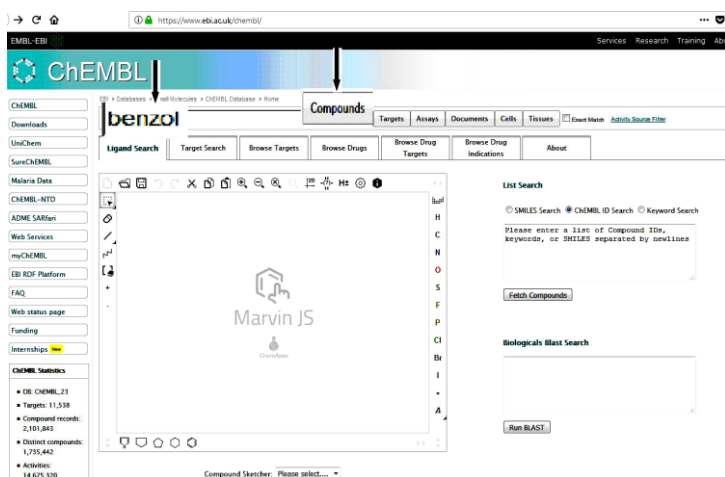
## ПРАКТИКУМ

### Завдання 1




У базі даних ChEMBL знайдіть хімічні сполуки, які містять у структурі бензолне кільце. Скільки таких записів?

### Рекомендація

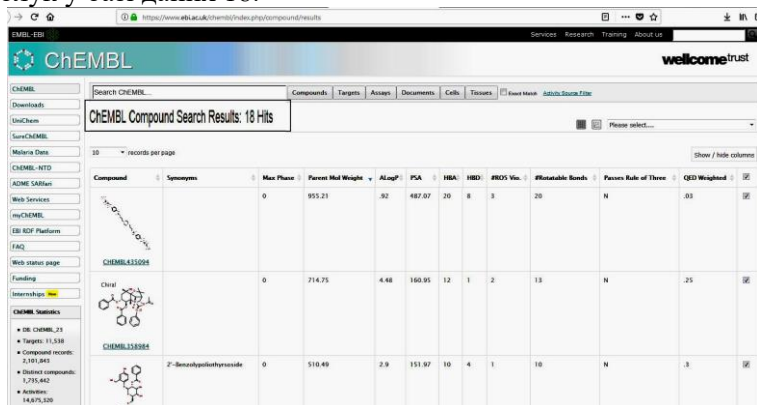
1. Здійснюємо вхід на вер-ресурс ChEMBL (<https://www.ebi.ac.uk/chembl>). У полі пошуку вводимо запит для бензолу (benzol) та справа натискаємо «Compounds».






ChEMBL Compound Search Results: 18 Hits

Compound	Synonym	Max Phase	Parent Mol Weight	AllogP	PSA	HBA	HBD	#R05 Vics	#Rotatable Bonds	Parent Rule of Three	QED Weighted
	CHEMBL331066	0	955.23	82	487.07	20	8	3	20	N	0.03
	Chiral	0	714.75	4.48	160.95	12	1	2	13	N	.25
	2-Benzoylphenylhydrazide	0	510.49	2.9	351.97	10	4	1	10	N	.8

2. Отримуємо результат пошуку, з якого видно, що таких сполук у базі даних 18.



ChEMBL Compound Search Results: 18 Hits

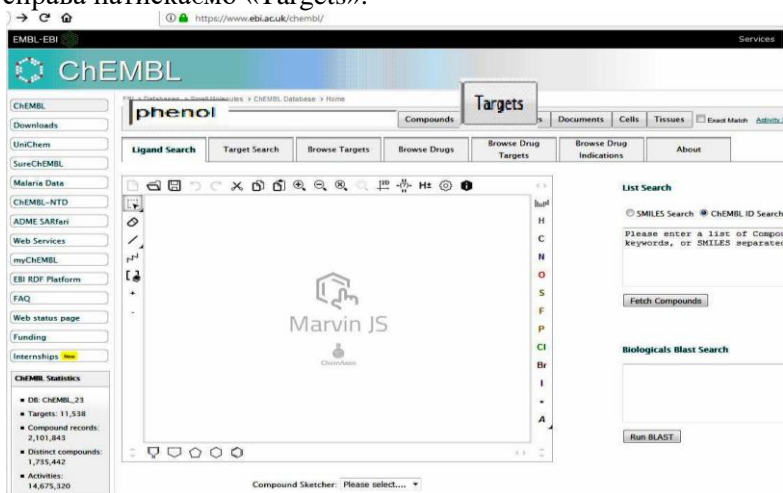
Compound	Synonym	Max Phase	Parent Mol Weight	AllogP	PSA	HBA	HBD	#R05 Vics	#Rotatable Bonds	Parent Rule of Three	QED Weighted
	CHEMBL331066	0	955.23	82	487.07	20	8	3	20	N	0.03
	Chiral	0	714.75	4.48	160.95	12	1	2	13	N	.25
	2-Benzoylphenylhydrazide	0	510.49	2.9	351.97	10	4	1	10	N	.8

## Завдання 2

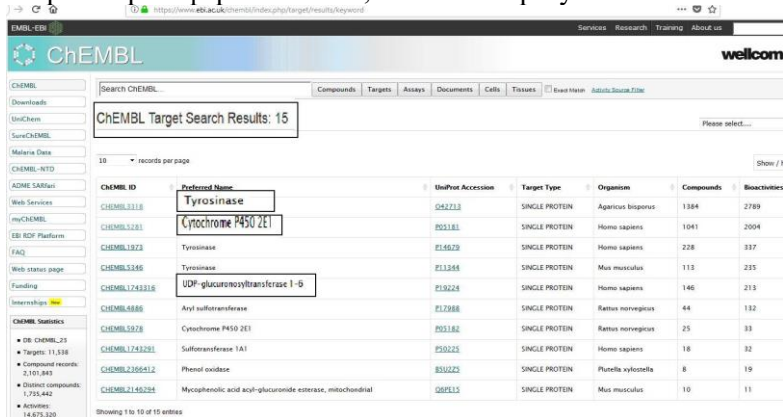
У базі даних ChEMBL знайдіть біологічні мішені, молекуло-лігандом яких виступає фенол.

### Рекомендація

1. Здійснюємо вхід на веб-ресурс ChEMBL (<https://www.ebi.ac.uk/chembl>). У полі пошуку вводимо запит для фенолу (phenol) та справа натискаємо «Targets».



2. Отримуємо результат пошуку, з якого видно, що мішенями дії фенолу є тирозиназа, цитохром P450 2E1, UDP-глюкоронілтрансфераза та інші, загалом 15 результатів.



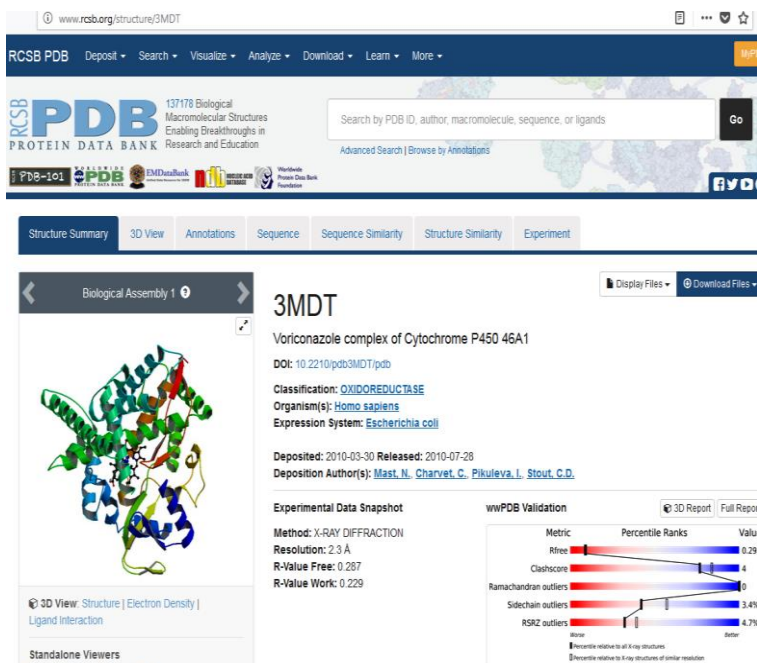
ChEMBL ID	Preferred Name	UniProt Accession	Target Type	Organism	Compounds	Bioactivities
CHEMBL3318	Tyrosinase	Q42713	SINGLE PROTEIN	Aperous biporus	1384	2769
CHEMBL3281	Cytochrome P450 2E1	P01161	SINGLE PROTEIN	Homo sapiens	1041	2004
CHEMBL1923	Tyrosinase	E15679	SINGLE PROTEIN	Homo sapiens	228	337
CHEMBL3346	Tyrosinase	P11354	SINGLE PROTEIN	Mus musculus	113	235
CHEMBL1743116	UDP-glucuronosyltransferase 1-B	P19324	SINGLE PROTEIN	Homo sapiens	146	213
CHEMBL4846	Aryl sulfotransferase	P17988	SINGLE PROTEIN	Rattus norvegicus	44	132
CHEMBL1928	Cytochrome P450 2E1	P01162	SINGLE PROTEIN	Rattus norvegicus	25	33
CHEMBL1743293	Sulfotransferase 1A1	P30225	SINGLE PROTEIN	Homo sapiens	18	32
CHEMBL2395412	Phenol oxidase	B3U2Z1	SINGLE PROTEIN	Pluccia xylostella	8	19
CHEMBL145294	Mycophenolic acid acyl-glucuronide esterase, mitochondrial	Q8P613	SINGLE PROTEIN	Mus musculus	10	11

### Завдання 3

Встановіть, чи бере участь залізо гему цитохрому P450 у зв'язуванні лікарського препарату воріконазолу.

#### Рекомендація

Здійснюємо вхід на веб-сервер PDB (<https://www.rcsb.org>) і знаходимо 3D-структуру комплексу цитохрому P450 з лікарським препаратом (cytochrome P450 drug). Вибираємо результат пошуку з ID 3MDT – комплекс воріконазолу (синтетичний протигрибковий препарат) з цитохромом P450.



3MDT

Voriconazole complex of Cytochrome P450 46A1

DOI: [10.2210/pdb3MDT/pdb](https://doi.org/10.2210/pdb3MDT/pdb)

Classification: [OXIDOREDUCTASE](#)

Organism(s): [Homo sapiens](#)

Expression System: [Escherichia coli](#)

Deposited: 2010-03-30 Released: 2010-07-28

Deposition Author(s): [Mast, N.](#), [Charvet, C.](#), [Pikuleva, I.](#), [Stout, C.D.](#)

**Experimental Data Snapshot**

Method: X-RAY DIFFRACTION

Resolution: 2.3 Å

R-Value Free: 0.287

R-Value Work: 0.229

**wwPDB Validation**

Metric	Percentile Ranks	Value
Rfree		0.291
Clashscore		4
Ramachandran outliers		0
Sidechain outliers		3.4%
RSRZ outliers		4.7%

2. На сторінці результату знаходимо інформацію про цитохром P450 та препарат, яка представлена у вигляді таблиці.

RCSB PDB Deposit Search Visualize Analyze Download Learn More

Small Molecules

Ligands (2 Unique)

ID	Chains	Name / Formula / InChI Key	2D Diagram & Interactions	3D Interactions
<p><b>HEM</b></p> <p><a href="#">Query on HEM</a></p> <p><a href="#">Download SDF File</a></p> <p><a href="#">Download CCD File</a></p>	A, B	<p><b>PROTOPORPHYRIN IX CONTAINING FE</b></p> <p><i>HEME</i></p> <p>C<sub>34</sub> H<sub>32</sub> Fe N<sub>4</sub> O<sub>4</sub></p> <p>KABFMBPWCXCRK-RGGAHWMASA-L</p>		<a href="#">Ligand Interaction</a>
<p><b>VOR</b></p> <p><a href="#">Query on VOR</a></p> <p><a href="#">Download SDF File</a></p> <p><a href="#">Download CCD File</a></p>	A, B	<p><b>Voriconazole</b></p> <p>(2<i>R</i>,3<i>S</i>)-2-(2,4-difluorophenyl)-3-(5-fluoropyrimidin-4-yl)-1-(1<i>H</i>-1,2,4-triazol-1-yl)butan-2-ol</p> <p>C<sub>16</sub> H<sub>14</sub> F<sub>3</sub> N<sub>5</sub> O</p> <p>BCEHBSKCWLPMDN-MGPLVRAMSA-N</p>		<a href="#">Ligand Interaction</a>

External Ligand Annotations

3. Зайшовши на опцію «VOR», можна дізнатися більше про структуру препарату. Видно, що хімічна сполука, яка лежить в основі препарату містить два шестичленні кільця та одне п'ятичленне.

www.rcsb.org/ligand/VOR

RCSB PDB 137178 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations

**VOR**

Voriconazole

VOR as a free ligand exists in 5 entries. Examples include: GAYG 4UYM 4ZE0

Find related ligands: Stereoisomers Similar ligands  
Chemical Structure Search

[View summary at Ligand Expo](#)

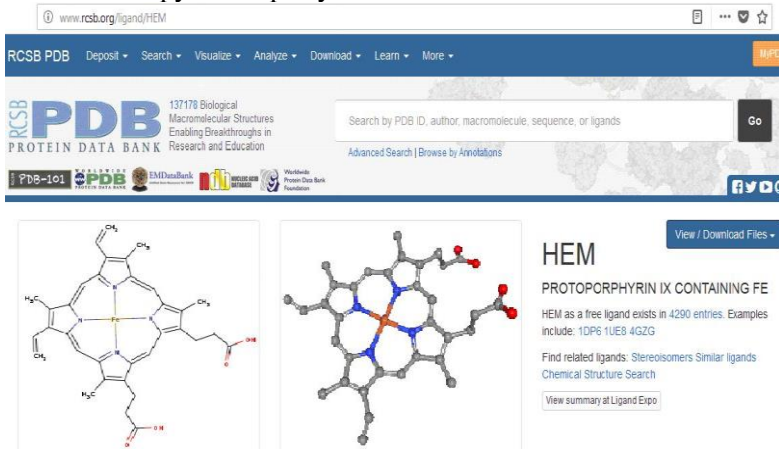
Rotate | Hydrogens | Labels

Chemical Component Summary

Chemical Details



4. Натиснувши на опцію «HEM», виявлено, що в структурі активного центру цитохрому P450 локалізований гем.



www.rcsb.org/ligand/HEM

RCSB PDB Deposit Search Visualize Analyze Download Learn More

RCSB PDB 137176 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

HEM

View / Download Files

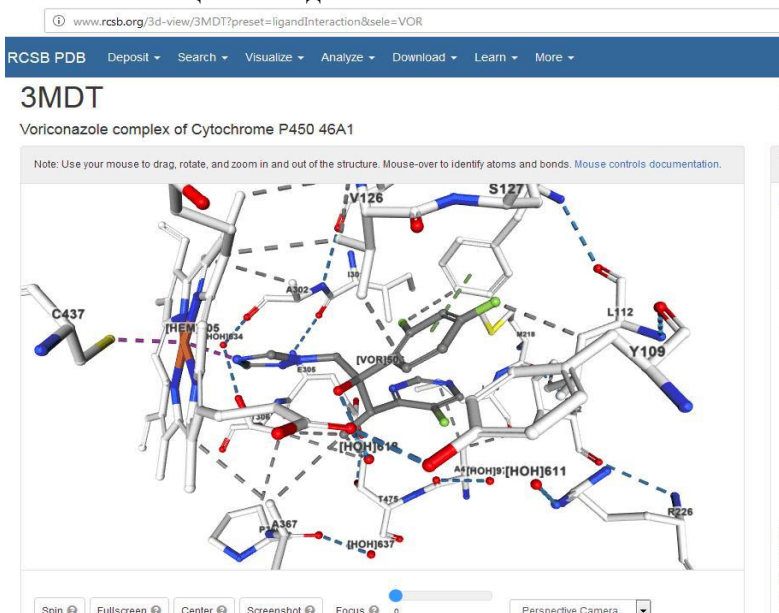
PROTOPORPHYRIN IX CONTAINING FE

HEM as a free ligand exists in 4290 entries. Examples include: 1DP6 1UE8 4GZG

Find related ligands: Stereoisomers Similar ligands  
Chemical Structure Search

View summary at Ligand Expo

5. Візуалізація 3D-структури комплексу цитохрому P450 з воріконазолом показала, що залізо гему цитохрому P450 взаємодіє з п'ятичленним кільцем хімічної сполуки, два інші шестичленні кільця взаємодіють з білковою частиною.



## 6. Збільшимо комплекс і детальніше розглянемо.

RCSB PDB Deposit Search Visualize Analyze Download Learn More

# 3MDT

Voriconazole complex of Cytochrome P450 46A1

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. [Mouse controls documentation.](#)

Structure View Electron Density Maps Ligand Viewer

Ligand View Documentation

The ligand pocket is only visible if the pocket opacity is set to greater than zero.

Ligand: IVOR1506

Pocket

Opacity: 0

Near Clipping: 0

Radius Clipping: 0

Color: Hydrophobicity

- Hydrogen Bonds (blue)
- Halogen Bonds (turquoise)
- Hydrophobic Contacts (grey)
- Pi Interactions (orange, green)
- Metal Interactions (purple)

Label

Show polymer display

Spin Fullscreen Center Screenshot Focus Perspective Camera

---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

### 1. Виберіть негативні модулятори фармакологічних мішеней:

- а) активатори;
- б) інгібітори;
- в) агоністи;
- г) антагоністи;
- д) деблокатори;
- е) блокатори.

2. Доповніть речення: *Методи добору лігандів як лідів для розробки ліків можуть бути об'єднані у а)..... та б) .....*

### 3. Виберіть етапи створення нового лікарського препарату, на яких можна застосовувати комп'ютерні методи:

- а) вибір хвороби, для лікування якої створюється лікарський засіб;
- б) вибір молекулярної мішені для лікарського засобу;
- в) знаходження базової структури для нового лікарського засобу;
- г) оптимізація структури та створення прототипу лікарського засобу;
- д) доклінічні дослідження;
- е) клінічні дослідження;
- ж) виробництво препарату.

### 4. До фармакологічних мішеней належать переважно .....

- а) молекули РНК;
- б) ферменти;
- в) рецептори;
- г) мембрани мітохондрій клітин;
- д) іонні канали.

5. Чи правильне твердження: *Основна мета ранніх стадій розробки ліків – встановлення однієї чи більше лідерних сполук?*

- а) так; б) ні

6. Стратегія пошуку біологічно активних молекул визначається тим, відомі чи ні тривимірні структури біоліганда і рецептора-мішені. При цьому цінніші знання .....

- а) структури рецептора;
- б) структури молекули-біоліганда.

7. До прямих методів, які використовують у комп'ютерному моделюванні ліків, належать:

- а) докінг;
- б) аналоговий дизайн;

- 
- 
- в) дизайн de novo;
  - г) фармакофорний скринінг.

**8. Вичленування окремих речовин із багатьох сполук, які проявляють однакову фармакологічну активність, називається ...**

- а) встановленням фармакофора;
- б) пошуком ліда;
- в) ідентифікацією фолда;
- г) вирівнюванням.

**9. Доповніть речення: *Найбільша частина відомих фармакологічних мішеней білкової природи (близько 50 % загальної кількості) належить до .....***

**10. Виберіть позитивні модулятори фармакологічних мішеней:**

- а) активатори;
- б) інгібітори;
- в) агоністи;
- г) антагоністи;
- д) деблокатори;
- е) блокатори.

**11. Виберіть правильну відповідь для заповнення пропуску: *Якщо структура мішені відома з експерименту, то можна застосувати молекулярне моделювання для .....* розробки ліків?**

- а) прямої;
- б) непрямої.

**12. Заповніть пропуски: ..... – це набір структурних ознак у молекулі, які розпізнаються біологічними рецепторами і відповідальні за біологічну активність молекули.**

**13. Вичленування функціональних груп (субструктур) у багатьох сполуках, які проявляють однакову фармакологічну активність або присідаються до однакових сайтів у білку, називається .....**

- а) ідентифікацією ортологів;
- б) рекомендацією ліда;
- в) встановленням фармакофора;
- г) ідентифікацією паралогів.

**14. Виберіть програми, які використовують для молекулярного докінгу під час розробки ліків:**

- 
- 
- a) Ramachandran Plot Explorer;
  - б) BLAST;
  - в) GOLD
  - г) T-Coffee;
  - д) AutoDock;
  - е) FlexX.

**15. Виберіть із переліку бази даних, у яких міститься інформація про велику кількість відомих хімічних сполук:**

- a) UniProt;
- б) SCOP;
- в) ChEMBL;
- г) GenBank;
- д) DDBJ.

**16. Виберіть положення, яке найповніше відображає мету застосування комп'ютерних методів у розробці ліків:**

- a) вивчати протеом організмів для створення ліків;
- б) створювати гіпотези про нові мішені дії ліків та ліганди, які взаємодіють із мішенями;
- в) створювати бази даних білків і нуклеїнових кислот.

**17. Чи можна засоби біоінформатики використовувати на етапах клінічних досліджень ліків:**

- a) так; б) ні.

**18. До непрямих методів комп'ютерного моделювання ліків належать .....**

- a) докінг;
- б) аналоговий дизайн;
- в) дизайн de novo;
- г) фармакофорний скринінг.

**19. Виберіть методи комп'ютерного моделювання, які використовують для виявлення та аналізу мішеней ліків.**

- a) ЯМР;
- б) докінг;
- в) кристалографія;
- г) фармакофорний скринінг.

**20. Чи правильне твердження: звичайні агоністи підвищують відповідь рецептора, зворотні агоністи знижують її?**

- a) так; б) ні.

---

---

## ТЕМА 6. БІОІНФОРМАТИКА В ІНТЕРПРЕТАЦІЇ МІКРОАРЕЙ-ЕКСПЕРИМЕНТУ І МАС-СПЕКТРОМЕТРІЇ

**Мета:** з'ясувати роль біоінформатики у збереженні даних та розробленні відповідних програм при мікроарей-аналізі та мас-спектрометрії. Навчитися користуватися веб-вправами та програмами, які допомагають аналізувати результати мікроарей-експерименту і мас-спектрометрії.

### ТЕОРЕТИЧНА ЧАСТИНА

#### 1. Мікроарей-біоінформатика

Для розуміння поширення та взаємодії білків у часі та просторі в клітині та організмі нині використовують високотехнологічні методи, зокрема мікроарей-аналіз та мас-спектрометрію, які допомагають сформувати цілісну картину наявності та роботи білків у живих системах, а також взаємозв'язок генів і білкових варіацій у популяціях.

Мікроарей-аналіз – новий напрям біологічних досліджень, орієнтований на інтегральне вивчення експресії генома. Його основа – широкий комплекс методів, які сприяють виділенню, детектуванню, ідентифікації та кількісному аналізу експресії десятків тисяч генів, із котрих складається транскриптом клітини. Важливу роль у встановленні сучасних мікроарей-технологій відіграло завершення проекту «Геном людини» та супутніх проектів із модельних організмів. Наявність розшифрованих геномів значно полегшила процедуру пошуку й анотації нових генів (як лабораторними, так і біоінформатичними методами), що зумовило удосконалення технологій виробництва мікроарейів.

Нині мікроарей-технології широко використовуються в аналізі диференційної генної експресії, детекції однонуклеотидного поліморфізму, філогенетичному аналізі, ідентифікації маркерів пухлин та розробці фармацевтичних препаратів.

Мікроарей являє собою скляний чи полімерний слайд, на який вибірково наносяться молекули ДНК (або білка).

ДНК мікроарейі аналізують:

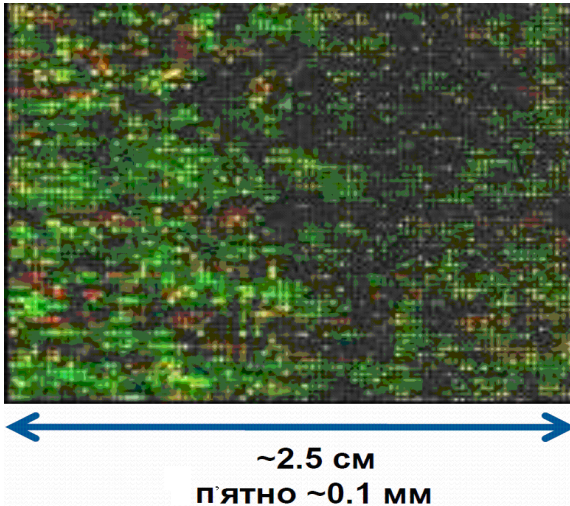
- мРНК у клітині та відповідний патерн експресії білків;

- геномні ДНК для ідентифікації відсутніх чи мутованих генів.

Для перевірки дослідного зразка на наявність багатьох послідовностей використовують ДНК-мікроареї або ДНК-чипи. Основна ідея полягає у тому, щоб встановити, чи нуклеотид із відомою послідовністю може приєднати комплементарно інший олігонуклеотид (один до одного) чи олігонуклеотиди (кілька до одного) зі зразка.

Типовий мікроарей-експеримент починається з гібридизації до мішеней на мікроарей-поверхні мічених флуоресцентними барвниками молекул ДНК (проб), одержаних зворотною транскрипцією та полімеразною ланцюговою реакцією (ЗТ-ПЛР), виділеної із клітин мРНК.

Після взаємодії арея зі зразком, кожний компонент, який зв'язався, приєднує флуорисцентну мітку. Оскільки послідовність нуклеотидів арея заздалегідь відома, встановлюють розміщення мітки в ареї. Результати мікроарей-експерименту виглядають, як кольорове зображення, де інтенсивність забарвлення відображає гібридизацію, яка відбулася. Два набори проб приєднують червоні та зелені флуорофори. Якщо гібридується одна проба, то мітка виявляється червоною або зеленою. У разі, якщо гібридизуються дві проби, мітка набуває жовтого кольору.



*Рис. 40. Результати мікроарей-аналізу*

---

На наступному етапі використовуються сканери та фотоімеджери для детекції інтенсивності флуоресценції кожної плями. Інтенсивність плями можна інтерпретувати як рівень генної експресії.

Перевагами даного методу є можливість водночас встановлювати рівень експресії тисяч і навіть десятків тисяч генів в одному зразку. Мікоарей-аналіз дає змогу також виявляти кількісний рівень експресії у зразку. Дані мікоарей-аналізу вказують лише на відносний рівень експресії і тому вважаються семіінформативними. Окрім того, рівень експресії мРНК не завжди відображає реальний рівень експресії білків.

Біоінформатика виділяє центральну роль в інтерпретації мікоарей-експерименту.

Виділяють три основні стадії в аналізі результатів досліджень мікоареїв:

- ідентифікація плям на сканованому мікоарей-зображенні;
- трансформація та нормалізація матриць ознак;
- аналіз матриць генної експресії.

За останні роки мікоарей-біоінформатика сформувалася як окрема галузь, відмінна від структурної та лінійної біоінформатики. Ще на ранніх етапах розвитку мікоареїв проблема менеджменту та стандартизації мікоарей-даних привернула до себе увагу в біоінформатиці. Велика різноманітність мікоарей-платформ, технік синтезу та мічення проб, умов гібридизації та методів постановки експерименту ускладнює створення уніфікованих баз даних для всіх типів мікоареїв.

Файловий формат мікоарей-даних набагато складніший, ніж формат послідовностей ДНК та білків. Саме тому досі немає архіву мікоарей-даних, на відміну від архіву бази даних білкових та нуклеотидних послідовностей. Сучасні мікоарей-технології допомагають порівнювати дані, одержані під час мікоарей-експериментів, а отже, наявність такого архіву сприяла б набагато кращому зрозумінню клітинних процесів, які контролюються на рівні транскрипції. Для розроблення уніфікованого формату мікоарей-даних створено товариство MGED (The Microarray Gene Expression Data Society). Найбільші мікоарей-бази даних такі:

**GEO** (<http://www.ncbi.nlm.nih.gov/geo/>);



---

---

*SMD* (<http://genome-www.stanford.edu/microarray>);

*ArrayExpress* (<http://www.ebi.ac.uk/arrayexpress>).

Отже, мікроарей-інформатика, створена на межі теорії алгоритмів, штучного інтелекту та параметричної статистики, призначена для розв'язання складних проблем біології.

## 2. Біоінформатика і мас-спектрометрія

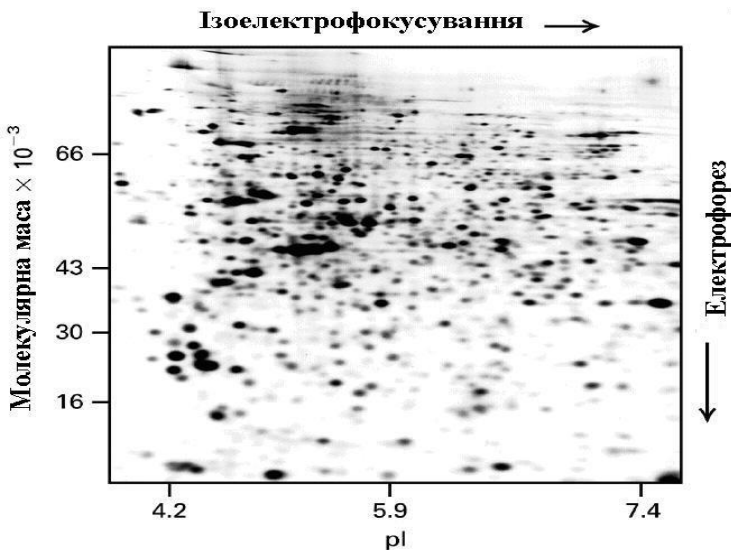
Мас-спектрометрія – фізичний метод, який дає змогу характеризувати молекули, визначаючи маси їхніх іонів. Застосування цього методу в біології допомагає:

- швидко ідентифікувати компоненти складної суміші білків;
- секвенувати білки та нуклеїнові кислоти, зокрема виконувати аналіз популяцій на наявність генетичних варіацій;
- аналізувати посттрансляційні модифікації чи заміни, які відбуваються у визначеній послідовності.

Для ідентифікації білків спочатку суміш білків піддають двовірному електрофорезу, після чого виділені білки розщеплюють трипсином для утворення фрагментів величиною 800 – 4000 амінокислотних залишків. Так із білкової молекули зі середньою довжиною 500 амінокислотних залишків отримують близько 50 криптичних фрагментів. Спектрометр визначає маси фрагментів з високою точністю, при цьому формується список мас фрагментів (відбитки пальців пептидних мас), який характеризує білок.

Двовірний (2D) електрофорез у поліакриламідному гелі (2D-PAGE) дає змогу розділити білки за двома параметрами: ізоелектричною точкою (pI) – величиною рН, при якій білок не має заряду і тому не рухається в електричному полі; молекулярною масою.

На рисунку 41 показаний приклад результату двовірного електрофорезу в поліакриламідному гелі. Кожна пляма відповідає визначеному білку. Білки у зразку розділенні за ізоелектричною точкою (по горизонталі) та молекулярною масою (по вертикалі).



***Рис. 41. Двовимірна електрофореграма білків***

Метод двовимірного електрофорезу також застосовують для аналізу диференційної експресії білка. За його допомогою можна визначити білки, які активуються або пригнічуються за дії різних лікарських препаратів; досліджувати білки, пов'язані з різними патологіями; слідкувати за змінами в експресії білка, які відбуваються протягом життєвого циклу клітин.

Результати експериментів 2D-PAGE зберігаються у базі даних 2D-PAGE і можуть бути знайдені на веб-сервері ресурсу ExPASy (<http://www.expasy.ch/ch2d>) – найбільшій базі даних щодо двовимірних гелів електрофорезу білків.

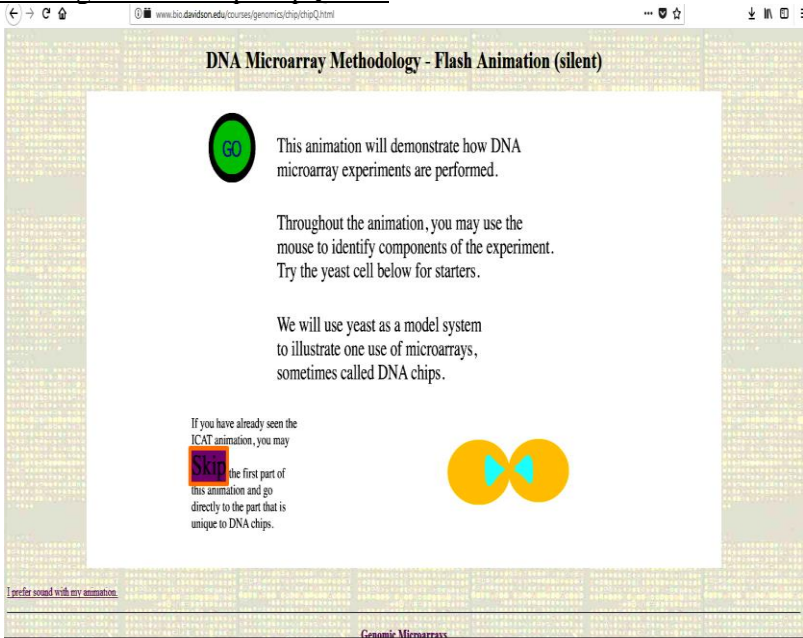
## ПРАКТИКУМ

### **Завдання 1**

Зайдіть на сайт Malcolm Campbell's у Davidson та перегляньте таку веб-вправу: DNA Microarray Methodology (а FLASH animation) – <http://www.bio.davidson.edu/courses/genomics/chip/chipQ.html>.

### **Рекомендація**

1. Здійснюємо вхід на веб-ресурс <http://www.bio.davidson.edu/courses/genomics/chip/chipQ.html> і натискаємо «Go».



### **Завдання 2**

Скільки записів про двовимірний електрофорез людських білків знаходиться на сайті Swiss-2DPAGE.

### **Рекомендація**

1. Здійснюємо вхід на веб-ресурс Swiss-2DPAGE – <http://www.expasy.org/ch2d> та в меню головного пошукового екрана вибираємо «by description».

**SWISS-2DPAGE**  
Two-dimensional polyacrylamide gel electrophoresis database

**SWISS-2DPAGE**  
SWISS-2DPAGE contains data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps. You can locate these proteins on the 2-D PAGE maps or display the region of a 2-D PAGE map where one expects to find a protein from UniProtKB/Swiss-Prot. [More details / References / Linking to SWISS-2DPAGE / Commercial uses]

Release 19.00, 23rd of May 2011, and updates up to the 9th of November 2011 (containing 1265 entries in 36 reference maps from human, mouse, Arabidopsis thaliana, Nicotystolium discolorum, Escherichia coli, Saccharomyces cerevisiae, and Staphylococcus aureus (K15)).

**Search by**  
 [accession number]  
 [description, ID or gene]  
 [author names]  
 [spot ID / serial number]  
 [identification methods]  
 [pI / Mw range]  
 [combined fields]

**Maps**  
 [experimental info]  
 [protein list]  
 [graphical interface]

Select Remote Interfaces  
 [All Interfaces]  
 World-2DPAGE Portal

**Access to SWISS-2DPAGE**  
 by description interface  
 [id in the ID, DE, GN and KW lines]  
 • by accession number (AC lines)  
 • by clicking on a spot: select one of our 2-D PAGE or SDS-PAGE reference maps, click on a spot and then get the corresponding information from the SWISS-2DPAGE database.  
 • by author (PK lines)  
 • by spot serial number (2D and 1D lines)  
 • by experimental pI/Mw range  
 • by experimental identification methods  
 • by full text search  
 • retrieve all the protein entries identified on a given reference map

**SWISS-2DPAGE documents**  
 • [Facts and statistics]  
 • User manual  
 • Release notes (September 26, 2006)  
 • FAQ (Frequently Asked Questions about SWISS-2DPAGE)  
 • Protocols:  
   ◦ Technical information about 2-D PAGE (PG's, silver staining, protocols, etc)  
 • Figure captions of SWISS-2DPAGE maps available from publications:  
   ◦ Human CSF, LLC, HEPG2, HEPG2SP, LIVER, LYMPHOMA, PLASMA, PLATELET, REK  
   UBET, U2OS, KIDNEY,  
   ◦ Dicyctostolium discolorum, Escherichia coli, Saccharomyces cerevisiae.

2. Переходимо за посиланням, вводимо ключове слово «human» і натискаємо «Execute query».

world-2dpage.epasy.org/swiss-2dpage/

SWISS-2DPAGE

Home | Contact  
World-2DPAGE Home

To improve security and privacy, we are moving our web pages and services from HTTP to HTTPS. To give users of web services time to transition to HTTPS, we will support separate HTTP and HTTPS services until the end of 2017. From January 2018 most HTTP traffic will be automatically redirected to HTTPS. [more...]

[View this page in https](#)

**SWISS-2DPAGE**  
SWISS-2DPAGE

**Search by**  
 [accession number]  
 [description, ID or gene]\*  
 [author names]  
 [spot ID / serial number]  
 [identification methods]  
 [pI / Mw range]  
 [combined fields]

**Maps**  
 [experimental info]  
 [protein list]  
 [graphical interface]

Select Remote Interfaces  
 [All Interfaces]  
 World-2DPAGE Portal  
 World-2DPAGE Repository

Exclude local DBs  
has only effect if a remote interface is selected

**SWISS-2DPAGE**  
Search by description (DE), entry name (ID), gene name (GN) or UniProtKB/Swiss-Prot keywords (KW)

Enter search keywords:

Limit to:  All fields  DE  ID  GN  KW

Include external UniProtKB data in search

Sort by:  Accession number  Protein ID  Gene name

Please enter a keyword. This may be any word or partial word appearing in the entry identifier (ID), the description (DE), the gene name (GN) or a UniProtKB/Swiss-Prot keyword (KW). For example, you may type spot/,human, or just spot, or APC1, HUMAN.

If you give more than one keyword, entries having any keyword will be listed. Please do NOT use any boolean operators (and, or, etc.); nor quotes (").

Copy

SWISS-2DPAGE

3. Из результатів пошуку видно, що кількість записів становить 415.

The screenshot shows the Swiss-2DPAGE website interface. At the top, there is a navigation bar with the logo and the text 'SWISS-2DPAGE'. Below this, a yellow banner contains a message about the transition from HTTP to HTTPS. The main content area features a search bar with the text 'human' entered. Below the search bar, a box displays 'Query Result: 415 matches'. A table of search results is shown, with columns for 'Accession number', 'ID', 'Description', 'Genes', 'Keywords', and 'Species'. Two results are highlighted in yellow:

Accession number	ID	Description	Genes	Keywords	Species
00029	CLIC1_HUMAN	ReName: Full-Chloride intracellular channel protein 1; ABName: Full-Chloride channel ABP; ABName: Full-Nuclear chloride ion channel Z7; Short+KCC27; ABName: Full-Regulatory nuclear chloride ion channel protein; Short+RNICC.	(Name=CLIC1; Synonyms=G6, NCC27)	3D-structure; Acetylation; Cell membrane; Chloride; Chloride channel; Complete proteome; Cytoplasm; Direct protein sequencing; Double bond; Glutathionylation; Ion transport; Ionic channel; Membrane, Nucleus; Phosphoprotein; Reference proteome; Transmembrane; Transmembrane helix; Transport; Voltage-gated channel	Homo sapiens (Human)
00048	ZNF53_HUMAN	ReName: Full-Zinc finger protein 533; ABName: Full-Zinc finger protein 786.	(Name=ZNF53)	3D-structure; Complete proteome; DNA-binding; Metal-binding; Nucleus; Polymorphism; Reference proteome; Transcription; Transcription regulation; Zinc; Zinc-finger	Homo sapiens (Human)

On the left side of the page, there are several navigation and search options, including 'Search by' (with fields for accession number, description, author names, etc.), 'Maps' (with links to experimental info, protein list, and graphical interface), and 'Select Remote Interfaces' (with checkboxes for 'All interfaces', 'World-2DPAGE Portal', and 'World-2DPAGE Repository').

### Завдання 3

На сайті Swiss-2DPAGE знайдіть фенілаланінгідроксилазу. Скільки посилань ви знайшли? Чи є тут інформація про людську фенілаланінгідроксилазу?

#### Рекомендація

1. Здійснюємо вхід на веб-ресурс Swiss-2DPAGE – <http://www.expasy.org/ch2d>, вибираємо «by description», в пошуковому полі вводимо запит для фенілаланінгідроксилази (phenylalanine hydroxylase) і натискаємо «Execute query». Виявлено 6 результатів пошуку, серед яких є інформація про фенілаланінгідроксилазу людини.

← ↻ 🏠 world-2dpagexpasy.org/swiss-2dpagel

Searching in "SWISS-2DPAGE" for entries matching any of the keywords:  
**phenylalanine hydroxylase**  
 in their description (DE), entry name (ES), gene names (GN) or UniProtKB Swiss Prot keywords (KW).

**Query Result: 6 matches**

Accession number	ID	Description	Genes	Keywords	Species
P06169	POD1_YEAST	RecName: Full-Pyruvate decarboxylase isozyme 1; EC=4.1.1.; EC=4.1.1.1;	(Name=POD1; OrderedLocustNames=YLR04MC; ORFNames=P2104)	3D-structure; Acetylation; Alkosteric enzyme; Branched-chain amino acid catabolism; Complete proteome; Cytoplasm; Decarboxylase; Direct protein sequencing; Lyase; Magnesium; Metal-binding; Nucleus; Phenylalanine catabolism; Phosphoprotein; Reference proteome; Thiamine pyrophosphate; Tryptophan catabolism	Saccharomyces cerevisiae (Baker's yeast)
P07227	POD1_HUMAN	RecName: Full-Protein disulfide-isomerase; Short=PODI; EC=5.3.4.1; AllName: Full-Cellular thymidyl hormone-binding protein; AllName: Full-Poly(4-hydroxy)ase subunit beta; AllName: Full-p25; Flags: Precursor;	(Name=PHB; Synonyms=ERBA2L; POD1; PODB)	3D-structure; Cell membrane; Disaperone; Complete proteome; Direct prote i sequencing; Disulfide bond; Endoplasmic reticulum; Isomerase; Membrane; Redox; ctnr center; Reference proteome; Reps; Signal	Homo sapiens (Human)
P07355	SYFB_ECOLI	RecName: Full-Phenylalanyl-tRNA synthetase beta chain; EC=6.1.1.20; AllName: Full-Phenylalanyl-tRNA ligase beta chain; Short=PhRS;	(Name=phT; OrderedLocustNames=bt1713; JW1703)	3D-structure; Aminocacyl-tRNA synthetase; ATP-binding; Complete proteome; Cytoplasm; Ligase; Magnesium; Metal-binding; Nucleotide-binding; Protein biosynthesis; Reference proteome; tRNA-binding; tRNA-binding	Escherichia coli
P08312	SYFA_ECOLI	RecName: Full-Phenylalanyl-tRNA synthetase alpha chain; EC=6.1.1.20; AllName: Full-Phenylalanyl-tRNA ligase alpha chain; Short=PhRS;	(Name=phS; OrderedLocustNames=bt1714; JW5277)	Aminocacyl-tRNA synthetase; ATP-binding; Complete proteome; Cytoplasm; Ligase; Magnesium; Metal-binding; Nucleotide-binding; Protein biosynthesis; Reference proteome	Escherichia coli
P08103	POA1_MOUSE	RecName: Full-Protein disulfide-isomerase; Short=PODI; EC=5.3.4.1; AllName: Full-Cellular thymidyl hormone-binding protein; AllName: Full-Endoplasmic reticulum resident protein 59; Short=ER protein 59; Short=ERp59; AllName: Full-Poly(4-hydroxy)ase subunit beta; AllName: Full-p25; Flags: Precursor;	(Name=Phib; Synonyms=Pda1)	Cell membrane; Chaperone; Complete proteome; Disulfide bond; Endoplasmic reticulum; Isomerase; Membrane; Redox active center; Reference proteome; Repeat; Signal	Mus musculus (Mouse)
P35505	FAAA_MOUSE	RecName: Full-Fumarylacetoacetase; Short=FAA; EC=3.1.1.2; AllName: Full-Beta-diolonase; AllName: Full-Fumarylacetoacetate hydrolase;	(Name=Fat)	3D-structure; Acetylation; Calcium; Complete proteome; Hydrolyase; Magnesium; Metal-binding; Phenylalanine catabolism; Reference proteome; Tyrosine catabolism	Mus musculus (Mouse)

**Left sidebar:**

**SWISS-2DPAGE**

**Search by**

- [accession number]
- [description\_id\_or\_gene] \*
- [author names]
- [spot ID / serial number]
- [identification methods]
- [pI / Mw range]
- [combined fields]

**Maps**

- [experimental info]
- [protein list]
- [graphical interface]

**Select Remote Interfaces**

- [All interfaces]
- World-2DPAGE Portal
- World-2DPAGE Repository
- Exclude local DBs  
has any effect if a remote interface is selected

## Завдання 4

Знайдіть цитохром P450 печінки миші. Використовуючи інформацію сайту, дайте відповідь на запитання: Яка молекулярна маса (Mw) та ізоелектрична точка (pI) цитохрому P450 миші згідно з даними, отриманими 2D електрофорезом? Ознайомтеся з 2D-електрофореграмою.

## Рекомендація

1. Здійснюємо вхід на веб-ресурс Swiss-2DPAGE – <http://www.expasy.org/ch2d>, вибираємо «by description», у пошуковому полі вводимо запит для цитохрому P450 миші (mouse cytochrome P450) і натискаємо «Execute query».

world-2dpagexpasyorg/swiss-2dpagde

SWISS-2DPAGE

To improve security and privacy, we are moving our web pages and services from HTTP to HTTPS. To give users of web services time to transition to HTTPS, we will support separate HTTP and HTTPS services until the end of 2017. From January 2018 most HTTP traffic will be automatically redirected to HTTPS. [more...]

View this page in https

---

**SWISS-2DPAGE**

**Search by**

[accession number]  
 [description, ID or gene] \*  
 [author names]  
 [spot ID / serial number]  
 [identification methods]  
 [pI / Mw range]  
 [combined fields]

**Maps**

[experimental info]  
 [protein list]  
 [graphical interface]

Select Remote Interfaces

[All interfaces]   
 World-2DPAGE Portal   
 World-2DPAGE Repository

**SWISS-2DPAGE**

Search by description (DE), entry name (ID), gene name (GN) or UniProtKB/Swiss-Prot keywords (KW)

Enter search keywords:

Limit to:  All fields  DE  ID  GN  KW

Include external UniProtKB data in search

Sort by:  Accession number  Protein ID  Gene name

Please enter a keyword. This may be any word or partial word appearing in the entry identifier (ID), the description (DE), the gene names (GN) or a UniProtKB/Swiss-Prot keyword (KW). For example, you may apoa1\_human, or just apo, or APO or APOA1\_HUMAN.

If you give more than one keyword, entries having any keyword will be listed. Please do NOT use any boolean operators (and, or, etc.), nor quotes ("").

SWISS-2DPAGE

2. Из поданих електрофореграм вибираємо ту, яка відповідає печінці миші, де вказано, що значення:  $M_w=26047$ ;  $pI=6,23$

world-2dpagexpasyorg/swiss-2dpag/ac=PRDX6\_MOUSE

**BAT\_MOUSE** (Brown adipose tissue)  
*Mus musculus* (Mouse)  
 Tissue: Brown adipose tissue

**BAT\_MOUSE**

**MAP LOCATIONS:**

- SPOT 2D-001QDU: pI=6.19; Mw=26097

**MAPPING (identification):**  
 GEL MATCHING WITH WAT\_MOUSE [1].

**LIVER\_MOUSE** (Liver)  
*Mus musculus* (Mouse)  
 Tissue: Liver

**LIVER\_MOUSE**

**MAP LOCATIONS:**

- SPOT 2D-00149J: pI=6.26; Mw=26252 [identification data]
- SPOT 2D-00149S: pI=5.90; Mw=26197 [identification data]

**MAPPING (identification):**  
 Peptide mass fingerprinting [1].

**MUSCLE\_MOUSE** (Gastrocnemius muscle)  
*Mus musculus* (Mouse)  
 Tissue: Gastrocnemius

**MUSCLE\_MOUSE**

**MAP LOCATIONS:** pI=6.23; Mw=26047 [identification data]

- SPOT 2D-0017JD:

**MAPPING (identification):**  
 Peptide mass fingerprinting [1].

**WAT\_MOUSE** (White adipose tissue)  
*Mus musculus* (Mouse)  
 Tissue: White adipose tissue

**WAT\_MOUSE**

**MAP LOCATIONS:**

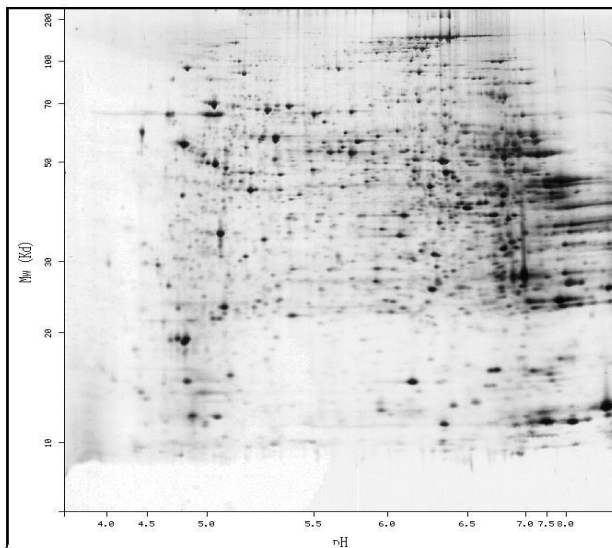
- SPOT 2D-001C17: pI=6.16; Mw=28186 [identification data]

**MAPPING (identification):**

### 3. Заходимо на електрофореграму.

← → ↻ 🏠 [world-2dpage.expasy.org/swiss-2dpage/viewer&map=LIVER\\_MOUSE&ac=O08](http://world-2dpage.expasy.org/swiss-2dpage/viewer&map=LIVER_MOUSE&ac=O08)

-Get more information by dragging your mouse pointer over any highlighted spot -Click on a highlighted spot to access all its ass [estimated location]





---

---

## ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ

**1. Чому досі немає світового архіву мікроарей-даних, на відміну від архіву бази даних білкових і нуклеотидних послідовностей?**

- а) недостатня кількість даних порівняно з даними послідовностей ДНК та білків;
- б) файловий формат мікроарей-даних набагато складніший, ніж формат послідовностей ДНК та білків;
- в) сучасні мікроарей-технології не дають змоги виконувати порівняння даних, одержаних під час мікроарей-експериментів

**2. Виберіть із переліку базу даних, яка є основною зі збереження інформації із двовимірних електрофорезів білків:**

- а) UniProt;
- б) ExPASy;
- в) BRENDA;
- г) GenBank;
- д) DDBJ;
- е) PDB.

**3. Фізичний метод, який допомагає характеризувати молекули, визначаючи маси їхніх іонів називається .....**

- а) мікроарей-аналіз;
- б) мас-спектрометрія.

**4. Чи можна за допомогою мікроарей-експерименту водночас встановлювати рівень експресії десятків тисяч генів в одному зразку?**

- а) так; б) ні.

**5. Виберіть найбільші мікроарей-бази даних:**

- а) UniProt;
- б) ExPASy;
- в) ArrayExpress;
- г) GenBank;
- д) GEO;
- е) PDB.

**6. Доповніть речення: Двовимірний електрофорез у поліакриламідному гелі (2D-PAGE), дає змогу розділити білки за двома параметрами: а) ..... та б) .....**

---

---

**7. Новий напрям біологічних досліджень, зорієнтований на інтегральне вивчення експресії генома, називається .....**

- а) мікроарей-аналізом;
- б) мас-спектрометрією.

**8. Чи правильне твердження: *Мікроарей-аналіз та мас-спектрометрія допомагають сформувати цілісну картину наявності та роботи білків у живих системах, а також встановити взаємозв'язок генів та білкових варіацій у популяціях?***

- а) так; б) ні.

**9. У якому методі використовують ДНК-чипи?**

- а) мікроарей-аналізі;
- б) мас-спектрометрії.

**10. Якого кольору набуває мітка під час мікроарей-аналізу у разі, коли гібридизується дві проби?**

- а) зеленого;
- б) жовтого;
- в) червоного.

---

---

## ВІДПОВІДІ ДО ТЕСТОВИХ ЗАПИТАНЬ ТА ЗАВДАНЬ

### ТЕМА 1. ВСТУП ДО БІОІНФОРМАТИКИ

- |                                      |                                        |
|--------------------------------------|----------------------------------------|
| 1) в;                                | 12) а;                                 |
| 2) а) біоінформатика послідовностей; | 13) в;                                 |
| б) структурна біоінформатика;        | 14) г;                                 |
| в) метаболічна біоінформатика;       | 15) а;                                 |
| 3) а, б, в, е;                       | 16) двовимірний гель-електрофорез;     |
| 4) б;                                | високоефективна рідинна хроматографія; |
| 5) базах даних (БД);                 | мас-спектрометрія;                     |
| 6) б;                                | 17) г;                                 |
| 7) б;                                | 18) д;                                 |
| 8) а;                                | 19) метаболом;                         |
| 9) а;                                | 20) б, в, д;                           |
| 10) транскриптоміка;                 |                                        |
| 11) б;                               |                                        |

### ТЕМА 2. БІОІНФОРМАЦІЙНІ БАЗИ ДАНИХ

- |                                 |                 |
|---------------------------------|-----------------|
| 1) б;                           | 15) б, г;       |
| 2) а) архівні бази;             | 16) г;          |
| б) бази даних, які курируються; | 17) а, г, д;    |
| в) похідні бази даних;          | 18) REM-TrEMBL; |
| 3) д;                           | 19) в;          |
| 4) б;                           | 20) г;          |
| 5) в, г, д;                     | 21) е;          |
| 6) г;                           | 22) в;          |
| 7) г;                           | 23) в;          |
| 8) д;                           | 24) б, в;       |
| 9) в;                           | 25) б;          |
| 10) г;                          | 26) PIR1;       |
| 11) д;                          | 27) а;          |
| 12) г;                          | 28) а;          |
| 13) д;                          | 29) а, б, д;    |
| 14) а;                          | 30) а;          |

### ТЕМА 3. ВИРІВНЮВАННЯ ПОСЛІДОВНОСТЕЙ І ФІЛОГЕНЕТИЧНІ ДЕРЕВА

- |          |       |
|----------|-------|
| 1) б, г; | 3) б; |
| 2) а;    | 4) б; |

- |                                                                                                                                                                                         |                                                                                               |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| 5) а) відстань Геммінга;<br>б) відстань Левенштейна.                                                                                                                                    | 16) філограма;<br>17) б;                                                                      |
| 6) вирівнювання послідовностей<br>гомологічних за всією довжиною;                                                                                                                       | 18) невикорінене;<br>19) а, д, е;                                                             |
| 7) а;                                                                                                                                                                                   | 20) множинним;                                                                                |
| 8) а;                                                                                                                                                                                   | 21) мультифуркаційного;                                                                       |
| 9) ортологи;                                                                                                                                                                            | 22) формат FASTA;                                                                             |
| 10) а, в, е;                                                                                                                                                                            | 23) б;                                                                                        |
| 11) а;                                                                                                                                                                                  | 24) вирівнювання, яке<br>перекривається;                                                      |
| 12) а, в, г;                                                                                                                                                                            | 25) б;                                                                                        |
| 13) мінімальна кількість операцій<br>редагування між двома ланцюга-<br>ми не обов'язково однакової<br>довжини, необхідних для того,<br>щоб зробити один ланцюг<br>ідентичний до іншого; | 26) б;<br>27) філогенетичне дере-<br>во, яке не містить<br>інформації про довжи-<br>ни гілок; |
| 14) б, в, г, д;                                                                                                                                                                         | 28) а;                                                                                        |
| 15) б;                                                                                                                                                                                  | 29) а, в, д;<br>30) а.                                                                        |

#### ТЕМА 4. МЕТОДИ ПЕРЕДБАЧЕННЯ СТРУКТУРИ ТА ФУНКЦІЙ БІЛКІВ

- |                                                                |                                                                                                                |
|----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| 1) в;                                                          | 16) функціональна<br>дивергенція;                                                                              |
| 2) б;                                                          | 17) в;                                                                                                         |
| 3. б;                                                          | 18) конформер;                                                                                                 |
| 4) в, г;                                                       | 19) а, г;                                                                                                      |
| 5. а, г;                                                       | 20) $г \rightarrow в \rightarrow б \rightarrow а$ ;                                                            |
| 6) в;                                                          | 21) ротамер;                                                                                                   |
| 7) конвергенція;                                               | 22) $г \rightarrow д \rightarrow в \rightarrow б \rightarrow$<br>$\rightarrow а \rightarrow е \rightarrow ж$ ; |
| 8) в;                                                          | 23) а, в;                                                                                                      |
| 9) в;                                                          | 24) в, г;                                                                                                      |
| 10) б;                                                         | 25) в;                                                                                                         |
| 11) б, в;                                                      | 26) б, д, е;                                                                                                   |
| 12) моделювання на основі шаблону;<br>моделювання без шаблону; | 27) б;                                                                                                         |
| 13) а;                                                         | 28) а;                                                                                                         |
| 14) б, в;                                                      | 29) фолдинг;                                                                                                   |
| 15) а, б;                                                      | 30) в.                                                                                                         |

---

---

## ТЕМА 5. БІОІНФОРМАТИКА У ПОШУКУ ТА РОЗРОБЦІ ЛІКІВ

- |                                  |                 |
|----------------------------------|-----------------|
| 1) б, г, е;                      | 11) а;          |
| 2) індуктивні та дедуктивні;     | 12) фармакофор; |
| 3) б, в, г;                      | 13) в;          |
| 4) б, в, д;                      | 14) в, д, е;    |
| 5) а;                            | 15) в;          |
| 6) а;                            | 16) б;          |
| 7) а, в;                         | 17) б;          |
| 8) б;                            | 18) б, г;       |
| 9) рецепторів клітинних мембран; | 19) б, г;       |
| 10) а, в, д;                     | 20) а;          |

## ТЕМА 6. БІОІНФОРМАТИКА В ІНТЕРПРЕТАЦІЇ МІКРОАРЕЙ- ЕКСПЕРИМЕНТУ І МАС-СПЕКТРОМЕТРІЇ

- 1) б;
- 2) б;
- 3) б;
- 4) а;
- 5) в, д;
- 6) ізоелектричною точкою та молекулярною масою;
- 7) а;
- 8) а;
- 9) а;
- 10) б.

---

---

## СПИСОК УМОВНИХ СКОРОЧЕНЬ

- DDBJ (the DNA Data Bank of Japan) – Японська база даних ДНК
- EBI (the European Bioinformatics Institute) – Європейський інститут біоінформатики
- EMBL (European Molecular Biology Laboratory) – Європейська лабораторія молекулярної біології
- ExPASy (Expert Protein Analysis System) – експертна система аналізу білків
- FASTA (Fast Alingment) – швидке вирівнювання
- NCBI (National Center for Biotechnology Information) – Національний центр біотехнологічної інформації
- PIR (Protein Information Resource) – білковий інформаційний ресурс
- RCSB (Research Collaboratory for Structural Bioinformatics) – Дослідницька колаборація структурної біоінформатики
- SIB (Swiss Institute of Bioinformatics) – Швейцарський інститут біоінформатики

---

---

## ЛІТЕРАТУРА

1. Federhen S. The NCBI Taxonomy database / S. Federhen // *Nucleic Acids Research*. – 2012. – Vol. 40. – P. 136-143.
2. Hoepfner M. A. NCBI Bookshelf: books and documents in life
3. Mehmood M. A. Use of bioinformatics tools in different spheres of life sciences // M. A. Mehmood, U. Sehar, N. Ahmad // *J Data Mining Genomics Proteomics*. – 2014. – Vol. 5, № 5. – P. 1-13.
4. NEWT, a new taxonomy portal / I. Q. H. Phan, S. F. Pilbout, W. Fleischmann, A. Bairoch // *Nucleic Acids Research*. – 2003. – Vol. 31, №13. – P. 3822-3823.
5. Perekhoda L.O. The application of pass-computer programme and molecular docking for the search of new anticonvulsants / L.O. Perekhoda // *Annals of Mechnikov Institute*. – 2014. – №4. – P. 55-60.
6. sciences and health care / M. A. Hoepfner // *Nucleic Acids Research*. – 2013. – Vol. 41. – P. 1251-1260.
7. Васильева Н. Ю. Биоинформатика. Множественное выравнивание. Филогенетические деревья : методическое пособие / Н. Ю. Васильева. – Одесса: Одесский национальный университет имени И. И. Мечникова, 2014. – 70 с.
8. Волинець Г.П. Методи структурної біоінформатики / Г.П. Волинець, В.Г. Бджола, С.М. Ярмолук // *Біотехнологія*. – 2010. – Т. 3, № 4. – С. 9-19.
9. Горобець С. В. Біоінформатика як основний інструмент нанобіотехнології та наномедицини / С. В. Горобець, О. Ю. Горобець, Д. О. Дереча // *Клиническая информатика и Телемедицина*. – 2008. – Т.4, вып. 5. – С. 41-49.
10. Ефимов А. В. Структурные деревья глобулярных белков / А. В. Ефимов // *Успехи биологической химии*. – 2004. – Т. 44. – С. 109-132.
11. Иванов А.С. Основные принципы конформационного разнообразия белков для медико-биологов / А.С. Иванов // *Биомедицинская химия*. – 2011. – Т. 57, вып. 5. – С. 31-60.
12. Интегральная платформа «От гена до прототипа лекарства» in silico и in vitro / [И. С. Иванов, А. В. Веселовский, А. В. Дубанов и др.] // *Рос. хим. журн.* – 2006. – Т. 1, № 2. – С. 18-35.
13. Івахно С.С. Мікроареї: огляд технологій та аналіз даних / С.С. Івахно, О.І. Корнелю // *Укр. біохім. журн.* – 2004. – Т. 76, № 2. – С. 5-19.
14. Конформационный анализ структурных мотивов типа  $\alpha$ - $\alpha$ -уголок в вычислительном эксперименте молекулярной динамики / В. Р. Руднев, А. Н. Панкратов, Л. И. Куликова // *Математическая биология и биоинформатика*. – 2014. – Т. 9, № 2. – С. 575-584.

- 
15. Ложко Д. М. Модель просторової структури AIMP1/P43 людини / Д. М. Ложко, О. І. Корнелюк // Scientific Journal «Science Rise:Biological Science». – 2016. – № 2. – С. 41-46.
  16. Огурцов А. Н. Методи біоінформаційного аналізу / А. Н. Огурцов. – Х. : НТУ "ХПИ", 2011. – 114 с.
  17. Огурцов А. Н. Основи біоінформатики : учеб. посібник / А. Н. Огурцов. – Х. : НТУ «ХПИ», 2013. – 400 с.
  18. Пятницький М. А. Предсказание взаимосвязанных белков методами сравнительной геномики in silico / М. А. Пятницький, А. В. Лисица, А. И. Арчаков // Биомед. химия. – 2009. – Т. 55, № 3. – С. 230-246.
  19. Створення репозитарію моделей просторової структури тубулінів як реалізація одного із стратегічних завдань віртуальної організації CSLabGrid / [О.М. Демчук, П.А. Карпов, С.П. Ожередов та ін.] // Фактори експериментальної еволюції організмів. – 2016. – Т. 19. – С. 208-211.
  20. Структурна біоінформатика в постгеномну еру / [К. О. Одиноць, С. С. Івахно, Д. Б. Ковальський та ін.] // Біополімери і клітина. – 2004. – Т.20, № 1-2. – С. 78-91.
  21. Шмараков І.О. Біоінформатика: навч.-метод. посібник. – Чернівці: Рута, 2008. – 76 с.

### Інтернет-ресурси

- AceDB – <http://www.acedb.org>  
BLAST – <https://blast.ncbi.nlm.nih.gov>  
Bookshelf – <http://www.ncbi.nlm.nih.gov/books>  
BRENDA – <http://brenda-enzymes.info>  
CATH – <http://www.cathdb.info>  
ClustalW – <http://www.ebi.ac.uk/Tools/clustalw>  
DALI – [http://ekhidna.biocenter.helsinki.fi/dali\\_server](http://ekhidna.biocenter.helsinki.fi/dali_server)  
DDBJ – <http://www.ddbj.nig.ac.jp>  
DPD – <http://www.biologie.uni-hamburg.de/b-online/ibc99/iopi/iopihome.html>  
EMBL – <http://www.ebi.ac.uk>  
EMBL-SVA – <http://www.ebi.ac.uk/embl/sva>  
EMP – <http://emp.mcs.anl.gov>  
Ensembl – <http://www.ebi.ac.uk/ensembl/index.html>  
ENZYME – <http://www.expasy.ch/enzyme>  
EPD – <http://epd.vitalit.ch/index.php>  
ExPASy – <https://www.expasy.org>  
FlyBase – <http://flybase.bio.indiana.edu>  
GAD – <http://geneticassociationdb.nih.gov>  
GenBank – <http://www.ncbi.nlm.nih.gov/Genbank>  
Genomes Server – <http://www.ebi.ac.uk/genomes>



---

---

GOBASE – <http://gobase.bcm.umontreal.ca>  
GPCRDb – <http://gpcrdb.org>  
HGMD – <http://archive.uwcm.ac.uk/uwcm/mg/hgmd.html>  
HIV-SD – <http://hiv-web.lanl.gov/content/hivdb/mainpage.html>  
IMGT – <http://www.imgt.org>  
ITIS – <http://www.itis.gov>  
Karyn's Genomes – <http://www.ebi.ac.uk/2can/genomes/index.html>  
KEGG – [www.kegg.jp](http://www.kegg.jp)  
MaizeGDB – <https://www.maizegdb.org/person?id=16635>  
MENDEL – <http://www.mendel.ac.uk>  
MEROPS – <https://www.ebi.ac.uk/merops>  
MGD – <http://www.informatics.jax.org>  
MIPS – <http://www.mips.biochem.mpg.de/proj/yeast>  
NDB – <http://ndbserver.rutgers.edu>  
NEWT – <http://www.ebi.ac.uk/newt>  
PDB – <https://www.rcsb.org>  
PGDIC – <http://www.nal.usda.gov/pgdic>  
PROSITE – <https://prosite.expasy.org>  
Proteome Analysis – <http://www.ebi.ac.uk/proteome/index.html>  
PubMed – <https://www.ncbi.nlm.nih.gov/pubmed>  
RasMol – <http://www.openrasmol.org>  
RDP – <http://rdp.cme.msu.edu>  
REBASE – <http://rebase.neb.com/rebase/rebase.html>  
RGD – <http://rgd.mcw.edu>  
SCOP – <http://scop.mrc-lmb.cam.ac.uk/scop>  
Scopus – <https://www.scopus.com>  
SGD – <http://www.stanford.edu/Saccharomyces>  
Species 2000 – <http://www.sp2000.org>  
SPGP – [http://www.sanger.ac.uk/Projects/S\\_pombe](http://www.sanger.ac.uk/Projects/S_pombe)  
SRS – <http://srs.ebi.ac.uk>  
STACK – <http://www.sanbi.ac.za/Dbases.htm>  
Taxonomy – <http://www.ncbi.nlm.nih.gov/Taxonomy>  
T-Coffee – <http://www.ebi.ac.uk/Tools/tcoffee/index.html>  
The Tree of Life project – <http://tolweb.org/tree/phylogeny.html>  
TRANSFAC – <http://genexplain.com/transfac>  
UniGene – <http://www.ncbi.nlm.nih.gov/UniGene>  
UniProtKB – <http://www.uniprot.org>  
WormBase – <http://www.wormbase.org>  
YPD – <https://www.yeastgenome.org>

---

---

## ЗМІСТ

<b>ПЕРЕДМОВА</b>	3
<b>РОЗДІЛ 1. ІСТОРІЯ ТА ЗАСОБИ БІОІНФОРМАТИКИ</b>	4
<b>ТЕМА 1. ВСТУП ДО БІОІНФОРМАТИКИ</b>	4
1. Місце біоінформатики в ланцюгу постгеномних технологій.	4
2. Шляхи розвитку біоінформатики.	11
3. Мета, завдання та напрямки біоінформатики.	14
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	18
<b>ТЕМА 2. БІОІНФОРМАЦІЙНІ БАЗИ ДАНИХ</b>	21
1. Класифікація баз даних біологічної інформації.	21
2. Нуклеотидні бази даних.	24
3. Бази даних білків і амінокислотних послідовностей.	31
4. Бази даних метаболічних шляхів.	40
5. Таксономічні бази даних.	43
6. Бази даних наукової інформації.	46
ПРАКТИКУМ	50
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	66
<b>РОЗДІЛ 2. МЕТОДИ БІОІНФОРМАЦІЙНИХ ДОСЛІДЖЕНЬ</b>	72
<b>ТЕМА 3. ВИРІВНЮВАННЯ ПОСЛІДОВНОСТЕЙ І ФІЛОГЕНЕТИЧНИЙ АНАЛІЗ</b>	72
1. Завдання та види вирівнювання послідовностей.	72
2. Основні підходи парного та множинного вирівнювання послідовностей.	77
3. Програмне забезпечення вирівнювання послідовностей. Міра подібності послідовностей.	85
4. Типи та спосіб побудови філогенетичних дерев.	88
ПРАКТИКУМ	92
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	105
<b>ТЕМА 4. МЕТОДИ ПЕРЕДБАЧЕННЯ СТРУКТУРИ ТА ФУНКЦІЙ БІЛКІВ</b>	109
1. Складність проблеми і типи методів моделювання білків.	109
2. Передбачення вторинної структури білка.	111
3. Моделювання просторової структури білка.	117
4. Програми структурного вирівнювання білків. Передбачення функцій білків.	126
ПРАКТИКУМ	131
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	145

---

---

<b>ТЕМА 5. БІОІНФОРМАТИКА У ПОШУКУ ТА РОЗРОБЦІ ЛІКІВ</b>	150
1. Основні етапи пошуку та розробки ліків.	150
2. Основні стратегії комп'ютерного моделювання ліків.	153
3. Модернізація ключових етапів процесу розробки лікарських засобів.	158
ПРАКТИКУМ	161
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	167
<b>ТЕМА 6. БІОІНФОРМАТИКА В ІНТЕРПРЕТАЦІЇ МІКРОАРЕЙ-ЕКСПЕРИМЕНТУ ТА МАС-СПЕКТРОМЕТРІЇ</b>	170
1. Мікроарей-біоінформатика.	170
2. Біоінформатика і мас-спектрометрія.	173
ПРАКТИКУМ	175
ТЕСТОВІ ЗАПИТАННЯ ТА ЗАВДАННЯ	181
<b>ВІДПОВІДІ ДО ТЕСТОВИХ ЗАПИТАНЬ ТА ЗАВДАНЬ</b>	183
<b>СПИСОК УМОВНИХ СКОРОЧЕНЬ</b>	186
<b>ЛІТЕРАТУРА</b>	187

---

---

**Кеца Оксана Віталіївна**

**ОСНОВИ БІОІНФОРМАТИКИ**

*Навчально-методичний посібник*

Літературний редактор *Ряднова В. П.*  
Комп'ютерний набір та верстка *Кеца О. В.*

Формат 60/84. Бумага офсетная. Печать цифровая. Тираж 100 экз.  
Опечатано Ч.П. Озеров  
г. Харьков, ул. Университетская, 33/9  
Свидетельство о госрегистрации №818604 от 02.03.2000.