

## ЛЕКЦІЯ № 7

### ТЕМА: 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 структури білків та передбачення їхніх функцій.

#### План:

1. Експериментальні та теоретичні методи визначення просторової структури білків
2. Методи аналізу білків на рівні первинної структури
3. Передбачення вторинної структури білка
4. Моделювання просторової структури білка
5. Передбачення функцій білків

#### Література:

1. Кеца О. В. Основи біоінформатики: навч.-метод. посібник. Чернівці : Чернівецький нац. ун-т ім. Ю. Федьковича, 2018. 192 с.
2. Одинець К. О., Корнелюк О. І. Методи аналізу та моделювання просторової структури білків. *Наукові записки*. 2001. Т.19. Біологія та екологія. С. 7-17 URL: <https://ekmair.ukma.edu.ua/server/api/core/bitstreams/36d1e0e2-8c46-4af7-a125-e7cda0471cd7/content>

### 1. ЕКСПЕРИМЕНТАЛЬНІ ТА ТЕОРЕТИЧНІ МЕТОДИ ВИЗНАЧЕННЯ ПРОСТОРОВОЇ СТРУКТУРИ БІЛКІВ

Визначення просторової (тривимірної, 3D-) структури білків є необхідним етапом для встановлення взаємозв'язку між структурою та функцією білків. Значні успіхи у секвенуванні геномів організмів сприяли появі нових даних про амінокислотні послідовності різноманітних білків. З них лише приблизно для 10 % відомих АК-послідовностей доступні структурні дані, що становить понад 15 тис. експериментально визначених структур. Тому для аналізу більшості білків дедалі важливішими стають комп'ютерні методи передбачення їх просторової структури, які належать до методів «біології in silico»

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

#### **Експериментальні методи:**

Метод **рентгеноструктурного аналізу (РСА)**, який дозволяє визначати координати атомів білків з високою точністю.

При проведенні РСА білків отримують стабільні кристали високоочищеного білка з високою дифрагуючою здатністю.

#### Особливості методу:

- 1) Різні білки мають *індивідуальні* умови кристалізації.
- 2) Деякі білки можуть кристалізуватися тільки у вигляді протеолітичних фрагментів або не кристалізуються взагалі.
- 3) Для мембранних білків необхідно розробляти спеціальні методи для кристалізації.
- 4) В багатьох випадках метод РСА не є ефективним при вивченні білок-білкових та білок-нуклеїнових комплексів.

#### **ЯМР-спектроскопія**

Спектроскопія ядерного магнітного резонансу (ЯМР) є альтернативним методом визначення просторової структури білка. Перевагою методу ЯМР є можливість визначити структуру білків у розчині в нативних умовах. *Значний недолік* - обмеженість встановлення структури відносно великих білків (понад 20 кДа).

На основі експериментально визначених структур білків були розроблені певні емпіричні правила та алгоритми передбачення їх 3D-структури, які згодом були покладені в основу молекулярного моделювання просторової організації білків. Ці методи розвивалися спочатку як допоміжні при уточненні РСА- та ЯМР-структур, що дало змогу встановити загальні правила упаковки поліпептидних ланцюгів, а згодом розвинути новий підхід на основі конформаційного аналізу та мінімізації вільної енергії молекул білків.

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

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 100100 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:

- кількість можливих просторових конфігурацій білків дуже велика;
- 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.

## 2. МЕТОДИ АНАЛІЗУ БІЛКІВ НА РІВНІ ПЕРВИННОЇ СТРУКТУРИ

На першому етапі роботи потрібно виходячи з амінокислотної послідовності білка передбачити його загальні властивості:

- доменна організація,
- вторинна структура,
- консервативні та варіабельні ділянки тощо.

***Банки даних послідовностей.***

Інформація про визначені послідовності (первинні структури) біополімерів зберігається у спеціалізованих **банках даних**, які безперервно і автоматизовано поповнюються. Доступ до цих послідовностей можливий через пошук за ключовими словами або методом пошуку за гомологією. В записах банків даних наводяться додаткові відомості та різні перехресні посилання (наукові статті, інформація про вторинну структуру, різні аналізи, їх похибки тощо). Для генів і геномів це банки даних GenBank, EMBL, DDBJ та ін., а для білків — GenPept, TrEMBL, Swiss-Prot, OWL, NBRF-PIR та ін.

#### **Передбачення фізико-хімічних властивостей білків.**

Чисельні властивості білків (молекулярна маса, ізоелектрична точка (pI), оптичні властивості та ін.) можна розрахувати, виходячи з їх АК-послідовностей за допомогою досить простих програм: Proteomic Tools (<https://www.proteometools.org/>), ExPASy (<https://www.expasy.org/>).

#### **Пошук за гомологією.**

Важливим етапом комп'ютерного аналізу білка є пошук гомологічних АК-послідовностей серед усіх відомих послідовностей. Існують численні програми пошуку за гомологією, які можуть використовувати різні матриці та алгоритми вирівнювання, і враховують не тільки ідентичні, а й схожі за властивостями АК залишки. Найбільш вживаною є програма BLAST, подібними за призначенням є також програми FastA, Blitz. Варіант пошуку за гомологією, коли процес вирівнювання повторюється ітеративно з уточненням ролі найбільш консервативних залишків, дозволяє знаходити значно віддалені гомологи.

#### **Множинне вирівнювання.**

Гомологічні АК-послідовності мають бути порівняні між собою методом множинного вирівнювання. При цьому можна виявити окремі консервативні АК-залишки та їх скупчення (мотиви, сигнатури), які зберігаються в ході дивергентної еволюції генів цих білків. Аналіз групи гомологічних білків дозволяє виявити схожі ділянки поліпептидного ланцюга, які відповідають структурно консервативним елементам просторової структури (SCR), та варіабельні ділянки, що відповідають структурним елементам з різною структурою (SVR). Множинне вирівнювання можна зробити за допомогою серверів Clustal W, MultAlin, PredictProtein.

#### **Пошук мотивів та передбачення функцій білків.**

Білкові мотиви — відносно короткі схожі АК-послідовності, які зустрічаються у різних білках. Як правило, вони відповідають найбільш консервативним ділянкам активних центрів ферментів, особливо важливим залишкам для утворення потрібної просторової структури, сайтам модифікації (наприклад, фосфорилування або глікозилювання) та ін. Для пошуку структурних та функціональних мотивів використовують ProSite, PRINTS, Motif, NetPhos, NetOglyc та ін. До цієї ж групи програм можна приєднати також пошук сигнальних пептидів у секреторних білках (SignalP), трансмембранних ділянок (TopPred, DAS, TMHMM), передбачення сайтів протеолізу (PESTFind), локалізації в клітині (PSORT) та ін.

**Філогенетичний аналіз АК-послідовностей** призначений для аналізу ступеня їх схожості і можливого еволюційного походження, а також для їх наочної візуалізації у вигляді різноманітних дерев, діаграм тощо. Знання поширеності білкових доменів серед різних білків та організмів різних філогенетичних груп і наприклад, еубактерій, архебактерій і еукаріотів) дозволяє оцінити час їх виникнення, функціональну роль та важливість для існування організмів.

### **3. 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. 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.

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

#### **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.

#### **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.

#### **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),  $\psi$  (пси) і  $\omega$  (омега):

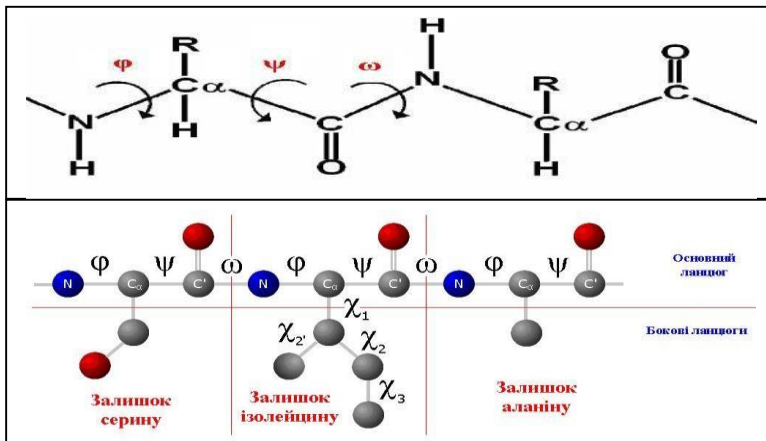


Рис. 1. Торсійні кути в поліпептидних ланцюгах

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).

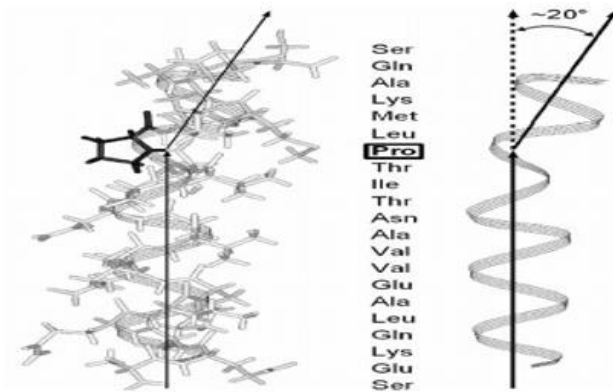


Рис. 2. Схема зображення згину у вторинній структурі білка

Для встановлення згинів у вторинній структурі білків будують профіль гідрофобності (рис. 3).

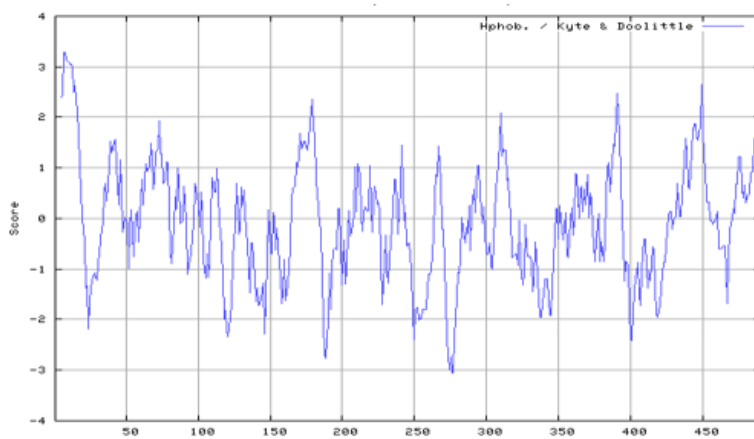


Рис. 3. Профіль гідрофобності

Значення нижче -2 вказують на згини у білковій молекулі.

#### 4. 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»*).

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. 4).

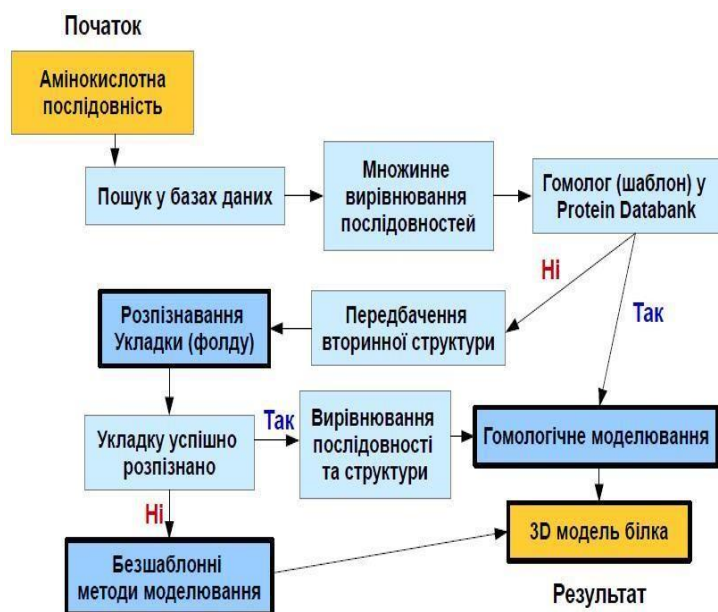


Рис. 4. Алгоритм вибору методу моделювання білків

### Моделювання на основі шаблону

#### Гомологічне моделювання

Суть методу полягає у побудові тривимірної моделі білка, структура якого невідома, на основі подібності його амінокислотної послідовності білка, структуру якого встановлено експериментально.

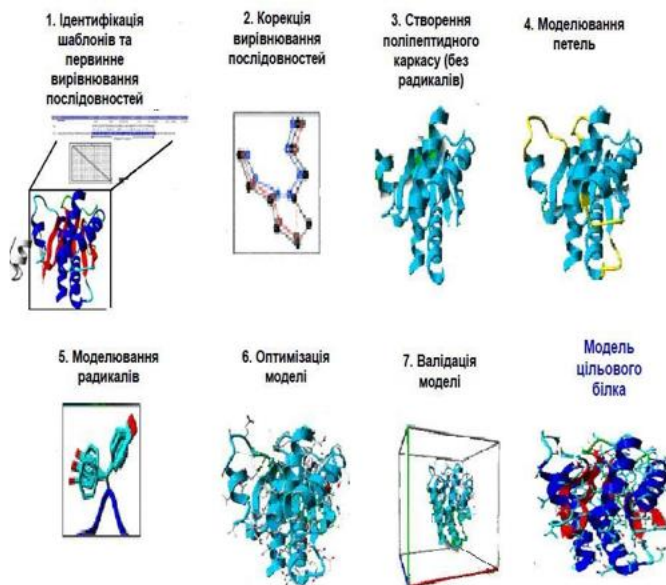


Рис. 5. Основні етапи гомологічного моделювання

#### 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 по чергово «протягується» через

кожний фолд. Далі відбувається сортування та відбір найкращих шаблонів. У кінцевому результаті створюється модель цільового білка.

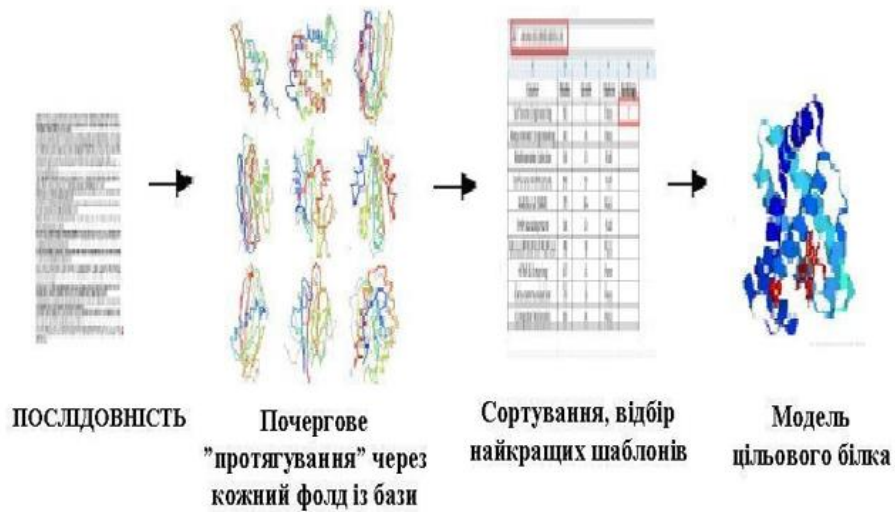


Рис. 6. Основні етапи моделювання білків методом «протягування»

### Моделювання без шаблону

#### 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.

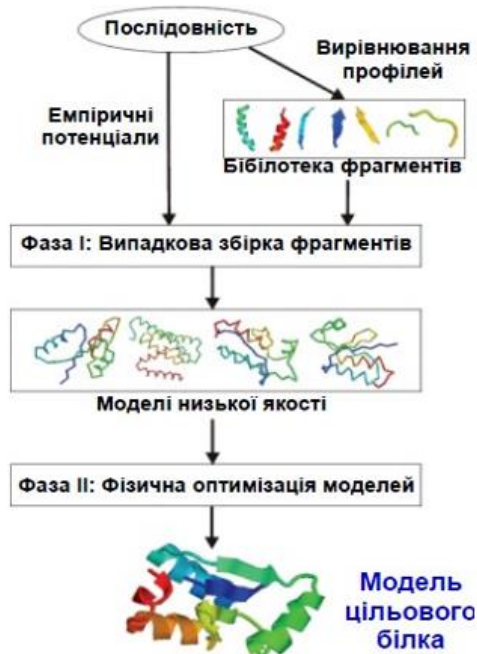


Рис. 7. Основні етапи моделювання білків de novo

## 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. 8. 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.

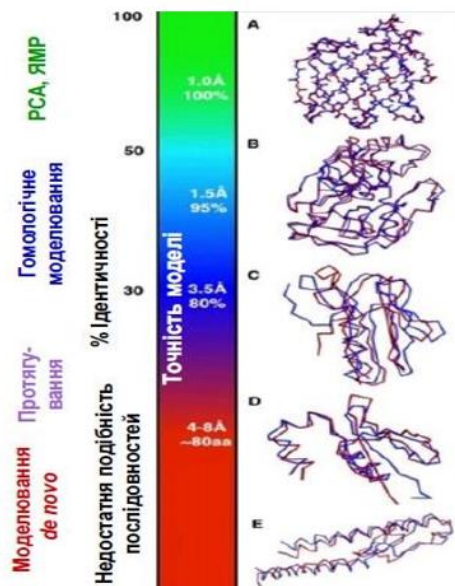
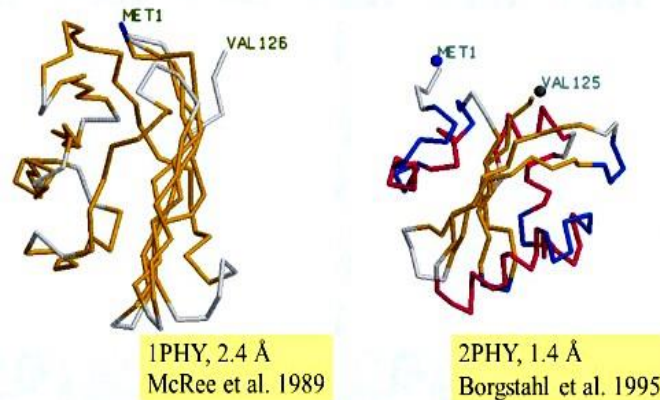


Рис. 9. Ефективність методів передбачення третинної структури білка



Часом у базі 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.



## 5. 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:

- 1) зберігати функцію і специфічність;
- 2) зберігати функцію, але змінювати специфічність;
- 3) змінювати функцію на подібну або ту ж саму, але в іншому метаболічному «контексті»;
- 4) 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 функцію з кількох подібних, або неофункціональності, 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 молюсків еволюціонував від різних предків. Так, гемоціанін членистоногих утворився з тирозинази, гемоціанін молюсків – із протеїнів. 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.