Кібернетика та комп'ютерні технології

УДК 519.272.2 DOI:10.34229/2707-451X.21.1.5

Б.О. БІЛЕЦЬКИЙ

ДОСЯГНЕННЯ У ВИЗНАЧЕННІ ПРОСТОРОВОЇ СТРУКТУРИ БІЛКІВ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Задачу визначення просторової структури білка відносять до найважливіших нерозв'язаних задач людства. Значних успіхів у цьому напрямку вдалося досягти, використовуючи біоінформатичні методи. Поєднання методів множинного вирівнювання, машинного навчання та оптимізації дозволили суттєво просунутись у визначенні білкових структур та досягти точності експериментальних методів. Задача визначення структури білків є настільки важливою та актуальною, що в дослідницьких колах говорять про першу нобелівську премію, яку може бути присуджено алгоритму машинного навчання.

Ключові слова: *просторова структура білка, AlphaFold, машинне навчання.*

Вступ. Задача визначення просторової структури білків – це одна з найважливіших нерозв'язаних задач людства. Життя на планеті Земля називають білковим, оскільки білкові молекули – рушії процесів життєдіяльності у живих організмах. Білки становлять близько 80 % сухої маси клітини та виконують різноманітні функції у живій клітині: вони формують клітинну мембрану та цитоскелет, які надають клітині розмір та форму; білки виконують транспортні функції доставки хімічних сполук до тих частин клітини, де ці сполуки потрібні; білки виробляють та перетворюють енергію, яка необхідна для підтримки процесів обміну речовин живої клітини; також білки копіюють та перетворюють генетичну інформацію. Взагалі, білки більше нагадують біомолекулярних нанороботів, принципи побудови яких не залишаються незрозумілими.

Кожен білок являє собою ланцюг амінокислотних залишків, пов'язаних між собою пептидними зв'язками. Своєї унікальної функції білок набуває після згортання амінокислотного ланцюга у просторову конфігурацію та утворення додаткових зв'язків (як правило водневих) між амінокислотними залишками ланцюга. Існує переконання, що амінокислотна послідовність білка, яка записана в генах молекули ДНК, повністю визначає просторову структуру білкової молекули і функцію білка.

Оскільки задача секвенування (читання) генетичної інформації з ДНК наразі вирішується відносно просто, то постає завдання швидкого визначення просторової структури білків за амінокислотними послідовностями. Існують експериментальні методи визначення просторової структури білкових молекул, такі як методи магнітно-ядерного резонансу та рентгеноструктурного аналізу. Такі методи є повільними та коштовними, наприклад, визначення просторової структури одного білка методом рентгеноструктурного аналізу коштує порядка $ 100000 та займає близько одного року роботи колективу вчених.

Станом на 2020 рік за допомогою експериментальних методів вдалося визначити структури для більш ніж 170 тис. білків, ці структури зберігаються у відкритих банках даних, таких як PDB [1]. Окрім того накопичено близько 200 млн. білків, структура яких залишається невідомою [2] (рис. 1). Накопичення білків з експериментально визначеною структурою а також дороговизна експериментальних методів визначення структури білків спонукають до розвитку альтернативних підходів до розв'язання цієї задачі. Однією з альтернатив є підхід на основі біоінформатики, де використовується комплекс методів машинного навчання, оптимізації, методів семплювання та інші.

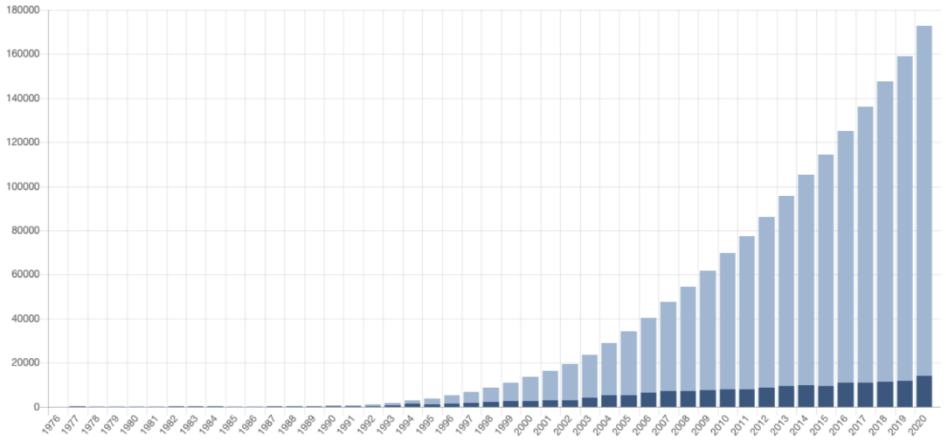

РИС. 1. Статистика з відкритої бази білкових структур PDB: щорічна загальна кількість білкових послідовностей (світлим) та щорічна кількість нових білкових структур (темним)

GDT та оцінки точності методів визначення білкових структур. З розвитком методів визначення білкових структур виникла необхідність оцінювати точність роботи таких методів. Точність визначення просторової структури білка вимірюється за допомогою методу GDT [3], який визначає міру схожості двох варіантів просторової структури білка з фіксованою амінокислотою послідовністю. Така міра схожості визначається як відсоток амінокислотних залишків, які потрапляють у заданий допустимий діапазон розбіжностей координат при накладанні двох просторових структур одна на одну.

CASP та порівняння методів визначення білкових структур. Існують ініціативи, які покликані порівнювати методи визначення білкових структур. Ці ініціативи носять характер відкритих змагань, де дослідницькі групи зі всього світу змагаються у визначенні структур білків. Для цього використовуються білки, структура яких отримана експериментально, але ще не опублікована. Деякі зі змагань, такі як EVA [4], орієнтовані на автоматичні методи, тоді як змагання CASP [5] порівнює точність експертних груп.

Для таких змагань виділяються негомологічні білки, що є окремою проблемою, адже гомологічні білки мають подібні амінокислотні послідовності та просторові структури що широко використовуються для побудови білкових структур за аналогією з гомологами. Існують ефективні методи пошуку гомологічних послідовностей у відкритих банках даних, так звані методи множинного вирівнювання, які також використовуються для виділення множин негомологічних білків.

Результати двох останніх змагань у рамках ініціативи CASP (які відбуваються кожні 2 роки) продемонстрували значні успіхи у передбачення просторової структури білків. Одна з дослідницьких груп представила метод AlphaFold, середня точність якого сягнула 87 GDT-балів. Варто зазначити, що точність експериментальних методів обмежена зверху 90 GDT-балами в силу специфіки експериментів з визначення структури білка.

Аналіз зображень та білкових структур за допомогою нейронних мереж. Принцип роботи методу AlphaFold грунтується на використанні штучних багатошарових нейронних мереж для визначення розподілів відстаней між залишками білкового ланцюга. При цьому структура представлення даних та їхньої обробки нагадують задачі аналізу зображень. Саме на задачах аналізу неструктурованих даних, зокрема зображень, вдалося досягти значних успіхів завдяки застосуванню багатошарових штучних нейронних мереж та глибинного навчання. Глибинне навчання нейронних мереж характеризується застосуванням множини шарів елементів, в яких відбуваються лінійні і нелінійні перетворення. При цьому шари відповідають за виділення представлень (або характеристик), які є більш зручним способом опису вхідних даних. Шари представлень утворюють ієрархічну структуру, тобто представлення вищого рівня виводяться з представлень нижчого рівня.

Наприклад, при застосуванні нейронних мереж для аналізу кольорових зображень розміром (1248x1248) пікселів, де кожен піксель представляється трьома компонентами градації червоного, зеленого та блакитного, вхідне зображення буде представлене шаром розмірності (1248x1248x3). В методі AlphaFold використовуються багатошарові нейронні мережі, де вхідні дані представляють собою матрицю характеристик пар амінокислотних залишків, кожен елемент такої матриці є ~700-вимірним вектором, тоді як при аналізі зображень кожен елемент аналогічної вхідної матриці кодувався 3-вимірним вектором.

З метою зменшення розмірності при аналізі зображень використовуються кілька інваріантних до зсуву (так званих конволюційних) блоків нейронних мереж, які являють собою багатошарову рамку зчитування, яка рухається по вхідному зображенню рис. 2.

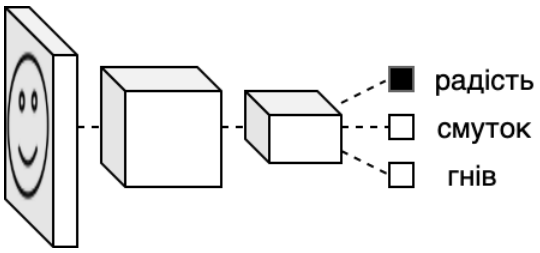

РИС. 2. Конволюційні нейронні мережі для аналізу для визначення емоцій

Через те, що багатошарові нейронні мережі дозволили досягти приголомшливих результатів при обробці зображень (і неструктурованих даних взагалі), з ними пов’язані певні недоліки, які часом запобігають їхньому застосуванню в галузях з високою ціною відмови. Неодноразово демонструвалися приклади, коли додавання до зображення збурень (інколи непомітних для людського ока) призводило до кардинально відмінних результатів роботи нейронними мережі [6]. Рис. 3 містить один з таких прикладів.

РИС. 3. Несподівані результати роботи нейронної мережі при додаванні до зображення непомітного шуму

Дослідження ефективності алгоритмів навчання залишається важливим напрямком в галузі штучних нейронних мереж. Важливим результатом у цьому напрямку є універсальна теорема апроксимації [7] та її різновиди. В ній йдеться про те, що широкі класи функцій можна наближувати за допомогою нейронних мереж. Існують різні варіанти цієї теореми для різних класів функцій та різних топологій нейронних мереж, однак такі результати носять переважно неконструктивний характер. Розглянемо одну з таких теорем – теорему Цибенка [8].

Теорема (Цибенка). Нехай $\phi: R \to R$ – активаційна функція нейронної мережі, яка має вигляд $\phi(\xi) = 1/(1+e^{-\xi})$, а також нехай $f: [0,1]^n \to R$ – неперервна функція.

Тоді

$$\forall \varepsilon > 0 \quad \exists N, w \in R^{N \times N}, \alpha \in R^N, \theta \in R^N, G(\cdot, w, \alpha, \theta): [0,1]^n \to R$$

такі що:

$$\left| G(x, w, \alpha, \theta) - f(x) \right| < \varepsilon, \ \forall x \in [0,1]^n,$$

де

$$G(x, w, \alpha, \theta) = \sum_{i=1}^{N} \alpha_i \phi(w_i^T x + \theta_i).$$

AlphaFold. Специфіка методу AlphaFold [9] полягає у використанні багатошарової нейронної мережі, яка навчена визначати дискретні розподіли визначальних для просторової структури білка величин, таких як торсіонні кути білкового ланцюга та попарні відстані між амінокислотними залишками білка. Згодом ці розподіли згладжуються та використовуються для побудови диференційованого білок-специфічного потенціалу, який, у свою чергу, застосовується для оптимізації енергії за допомогою градієнтного спуску та методів обмеженого семплювання.

На вхід методу потрапляє послідовність амінокислотних залишків білка $s = (s_1, ..., s_L)$ довільної довжини L. Метод повертає послідовність просторових координат залишків білкового ланцюга $c = (c_1, ..., c_L)$, де i-му залишку відповідає впорядкована трійка координат $c_i = (x_i, y_i, z_i) \in R^3$, $i = \overline{1, L}$.

Послідовність просторових координат $c = (c_1, ..., c_L)$ однозначно визначається послідовністю пар торсіонних кутів (ϕ, ψ), де $\phi = (\phi_1, ..., \phi_L)$ та $\psi = (\psi_1, ..., \psi_L)$, причому i-му амінокислотному залишку s_i білкового ланцюга відповідає пара торсіонних кутів (ϕ_i, ψ_i).

Просторова структура білка, тобто послідовність пар кутів (ϕ, ψ), визначається шляхом мінімізації енергії методом градієнтного спуску. Для цього для кожної білкової послідовності s будується диференційований білок-специфічний потенціал $V(\phi, \psi)$, який використовується для мінімізації енергії у процесі знаходження білкової структури. Білок-специфічний потенціал $V(\phi, \psi)$ має вигляд суми трьох складових:

$$V(\phi, \psi) = V_T(\phi, \psi) + V_D(\phi, \psi) + V_W(\phi, \psi),$$

де $V_T(\phi, \psi)$ – складова, яка відповідає за енергію торсіонних кутів, $V_D(\phi, \psi)$ – складова, яка пов'язана з енергією попарних відстаней між амінокислотними залишками, та $V_W(\phi, \psi)$ – складова, яка враховує сили Ван дер Ваальса. Складова $V_W(\phi, \psi)$, яка враховує сили Ван дер Ваальса, визначається за допомогою спеціального програмного комплекса Rosetta [10].

Складові $V_T(\phi,\psi)$ та $V_D(\phi,\psi)$, які описують, відповідно, внесок торсіонних кутів та попарних відстаней між залишками, визначаються за допомогою багатошарової нейронної мережі. На вхід нейронна мережа отримує матрицю $A=(a_{ij})_{L\times L}$ характеристик пар залишків. Кожній парі залишків (s_i,s_j) відповідає елемент $a_{ij}\in R^K$ матриці A, який є K-вимірним вектором, де K близько 700. В кожному такому векторі послідовно закодовані константи, індивідуальні характеристики i-го та j-го амінокислотних залишків (близько 100 індивідуальних характеристик для кожного залишку), а також парні характеристики залишків (s_i,s_j) (близько 500 парних характеристик). При цьому важливу роль відіграють характеристики, які визначаються за допомогою методів множинного вирівнювання HHblits [11] та PSI-BLAST [12] та містять еволюційну інформацію про залежність мутацій амінокислотних залишків у гомологічних білках.

Нейронна мережа, отримавши матрицю парних характеристик $A=(a_{ij})_{L\times L}$, повертає матрицю $D=(d_{ij})_{L\times L}$ дискретних розподілів відстаней між i-им та j-им залишком, де $d_{ij}(r)$ – оцінка ймовірності потрапляння відстані r між i-им та j-им залишками до одного з 64 напівінтервалів, на які розбитий відрізок 0-22 Å (контакту відповідає відстань < 8Å).

Окрім матриці розподілів D, нейронна мережа також повертає послідовність $T=(T_1,...,T_L)$ дискретних розподілів для пар торсіонних кутів, причому з i-им залишком білкового ланцюга пов'язаний розподіл $T_i(\phi_i,\psi_i)$, $i=\overline{1,L}$, який є оцінкою дискретного розподілу ймовірностей для i-ої пари торсіонних кутів (ϕ_i,ψ_i). Область значень кожного з торсіонних кутів розбивається на 36 напівінтервалів по 10°, таким чином кожна пара торсіонних кутів (ϕ_i,ψ_i) може приймати 1296 дискретних значень.

Дискретні розподіли D та T згладжуються та використовуються для побудови потенціалу $V(\phi,\psi)$ для мінімізації енергії за допомогою обмеженого семплювання. Семплювання називається обмеженим, оскільки популяція складається з 20 варіантів білкової структури. На першому кроці популяція генерується з розподілів торсіонних кутів T. Після чого кожна конфігурація з популяції проходить фазу оптимізації методом градієнтного спуску, а саме L-BFGS [13]. Нарешті популяція проходить штучний відбір і 10 % структур з вищим значенням енергії відсіюються. Решта 90 % структур піддаються мутаціям і потрапляють до наступного циклу оптимізації. До наступного циклу, замість відсіяних 10 %, потрапляють нові структури, згенеровані з розподілів торсіонних кутів T. Оптимізація білок-специфічного потенціалу $V(\phi,\psi)$ для кожного білкового ланцюга триває близько 5000 циклів.

Таким чином роботу методу AlphaFold (рис. 4) можна розбити на наступні кроки:

- на вхід подається амінокислотна послідовність білка;
- побудова матриці $A=(a_{ij})_{L\times L}$ характеристик пар залишків (s_i,s_j) за допомогою методів множинного вирівнювання;
- побудова розподілів D та T за допомогою багатошарової нейронної мережі;
- побудова потенціалу $V(\phi,\psi)$ за допомогою розподілів D та T та програмного комплексу Rosetta;
- мінімізація енергії за допомогою методу градієнтного спуску та обмеженого семплювання;
- просторова структура білка $c=(c_1,...,c_L)$ на виході.

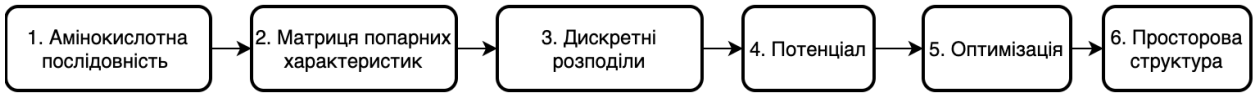

РИС. 4. Принцип роботи методу AlphaFold 2018

Структура нейронної мережі, яка визначає розподіли попарних відстаней показана на рис. 5. Зліва зображені вхідні дані розбиті на блоки. Перший шар розмірністю $L \times L \times 1$ кодує 1 глобальну змінну, наступні 2 блоки шарів кодують індивідуальні характеристики i-го та j-го залишків, за якими слідує блок шарів, які кодують характеристики пари залишків. Вхідні дані потрапляють до конволюційної (інваріантної до зсувів) нейронної мережі, яка відіграє роль рамки зчитування розміром 64×64 залишків. В глибину мережа складається з 220 блоків, кожен з яких містить 8 шарів.

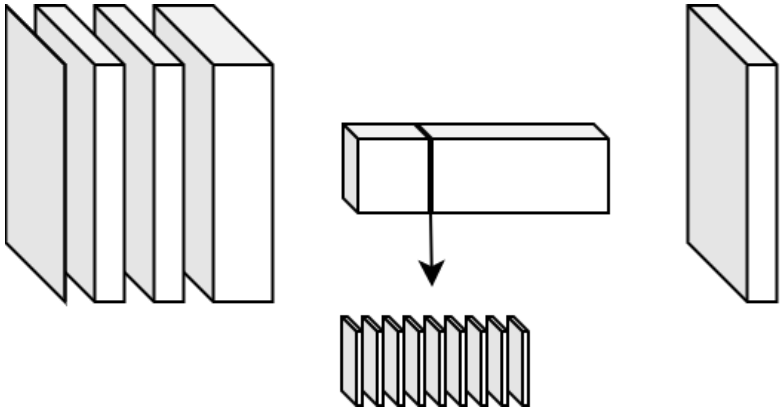

РИС. 5. Структура нейронної мережі AlphaFold

Для навчання нейронної мережі використовувалася вибірка з близько 30 тис. білкових структур з PDB, серед яких було обрано негомологічні домени за допомогою методу CATCH [14] з 35 % порогом гомологічності.

Висновки. В роботі розглянуто досягнення у галузі розпізнавання структури білків на основі методів машинного навчання. Зокрема, розглянуто принципи роботи програмного комплексу AlphaFold на основі поєднання цілої низки методів, таких як множинне вирівнювання, семплювання, оптимізація та інші. Важливим доробком колективу розробників AlphaFold є багатошарова нейронна мережа, яка будує оцінки дискретних розподілів попарних відстаней між амінокислотними залишками білкового ланцюга, а також оцінки дискретних розподілів значень пар торсіонних кутів білкового ланцюга. Іншим важливим внеском команди AlphaFold є використання білок-специфічного диференційованого потенціалу для пошуку структури з мінімальною енергією. Вдале поєднання різноманітних методів дозволяє значно підвищити точність розпізнавання просторової структури білків.

Список літератури

1. https://www.rcsb.org/stats/growth/growth-released-structures (звернення 10.02.2021)
2. https://www.uniprot.org/statistics/Swiss-Prot, https://www.ebi.ac.uk/uniprot/TrEMBLstats (звернення 10.02.2021)
3. Zemla A. LGA: A method for finding 3D similarities in protein structures. 2003. *Nucleic Acids Research*. **31** (13). P. 3370–3374. https://doi.org/10.1093/nar/gkg571
4. EVA: EValuation of Automatic protein structure prediction. http://pdg.cnb.uam.es/eva/doc/concept.html (звернення 10.02.2021)

5. Protein Structure Prediction Center. https://predictioncenter.org/index.cgi (звернення 10.02.2021)
6. Heaven D. Why deep-learning AIs are so easy to fool. *Nature*. 2019. **574**. P. 163–166. https://doi.org/10.1038/d41586-019-03013-5
7. Pinkus A. Approximation theory of the MLP model in neural networks. *Acta Numerica*. 1999. 8. P. 143–195. https://doi.org/10.1017/S0962492900002919
8. Cybenko G.V. Approximation by Superpositions of a Sigmoidal function. *Mathematics of Control Signals and Systems*. 1989. **2** (4). P. 303–314. https://doi.org/10.1007/BF02551274
9. Senior A.W., Evans R., Jumper, J. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*. 2020. **577**. P. 706–710. https://doi.org/10.1038/s41586-019-1923-7
10. https://www.rosettacommons.org/software (звернення 10.02.2021)
11. https://toolkit.tuebingen.mpg.de/tools/hhblits (звернення 10.02.2021)
12. https://blast.ncbi.nlm.nih.gov (звернення 10.02.2021)
13. Malouf R. A comparison of algorithms for maximum entropy parameter estimation. Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002). 2002. P. 49–55. https://doi.org/10.3115/1118853.1118871
14. https://www.cathdb.info/ (звернення 10.02.2021)

Одержано 10.02.2021

Білецький Борис Олександрович,
кандидат фізико-математичних наук,
старший науковий співробітник Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
borys.biletskyy@gmail.com

UDC 519.272.2

B.O. Biletskyy

Progress in Determination of Protein Spatial Structure Based on Machine Learning

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
Correspondence: borys.biletskyy@gmail.com

Introduction. The task of determining the spatial structure of proteins is one of the most important unsolved problems of mankind. Life on the planet Earth is called protein, because protein molecules are the drivers of life processes in living organisms. Proteins make up about 80% of the dry mass of the cell and coordinate the processes of metabolism. The functions of proteins are defined by its spatial structure.

The results of recent competitions in methods for determining protein structures have shown significant progress in this important area. One of the research groups presented the AlphaFold 2 method, the accuracy of which reached the accuracy of experimental methods.

Purpose of the article. The aim of the work is to consider and analyze the basic principles of the AlphaFold software package for determining the spatial structure of proteins.

Results. We consider the main stages in the process of recognizing the structure of a protein using the AlphaFold program complex. The stages and corresponding methods include: search for homologous proteins based on multiple alignment methods, construction of protein-specific differentiated potential using artificial neural networks and protein structure energy optimization using gradient descent and limited sampling. We discuss how combination of various bioinformatics techniques powered by data from open data sources can lead to significant improvements in accuracy of protein structure prediction. Special attention is paid to the use of artificial neural networks for building the smooth protein-specific potential and following energy minimization based on constructed potential.

Conclusions. The combination of a number of methods and the use of information from protein and genetic data banks allows us to make significant progress in solving the extremely important task of determining the structure of a protein.

Keywords: protein spatial structure, Machine Learning, AlphaFold.