ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ КОМП'ЮТЕРНОГО МОДЕЛЮВАННЯ ТРАНСКРИПЦІЙНОГО ФАКТОРА СТРЕСОСТІЙКОСТІ WRKY2 У TRITICUM AESTIVUM

Петровський А.П.[1], Дем'яненко І.В.[2]
[1] Custom PC Software, petrovskyyanton@gmail.com
[2]КПІ ім. Ігоря Сікорського, iryna.demjanenko@gmail.com

Abstract
This study conducts a comparative analysis of computational modelling methods, focusing on the transcription factor stress tolerance WRKY2 in wheat (Triticum aestivum). WRKY2 plays a vital role in stress response mechanisms in wheat. Utilizing advanced techniques such as machine learning (in particular, AlphaFold) and homology modelling, the efficacy of these methods in elucidating the structural and functional aspects of WRKY2 was compared. Our findings provide insights into the most suitable computational strategies for studying stress tolerance mechanisms mediated by WRKY2, contributing to the enhancement of crop resilience against environmental stresses.
Keywords: *protein modelling, machine learning, TaWRKY2, AlphaFold2.*

Вступ. В сучасному аграрному секторі зернові культури, особливо пшениця, мають надважливу роль для забезпечення глобальної продовольчої безпеки. В цьому контексті Україна, яка є великим виробником та експортером зернових, стикається з необхідністю підвищення врожайності та стійкості до стресових умов. Транскрипційні фактори, зокрема TaWRKY2 у пшениці, відіграють ключову роль у регуляції відповіді на стрес [1].

Розуміння молекулярних механізмів, що стоять за їх функціонуванням, є важливим для подальшого розвитку стресостійких сортів зернових культур. Одним з методів, який дасть змогу досягти поставленої мети, є встановлення його 3D структури. Для цього використовують як традиційні методи (рентгеноструктурний аналіз та ЯМР-спектрометрію), так і сучасні комп'ютеризовані методи. У сфері біології та біоінформатики вивчення структури та функцій білків має вирішальне значення.

Метою цієї роботи є визначення 3D структури транскрипційного фактору TaWRKY2 різними методами структурної геноміки та порівняння отриманих результатів.

Матеріали та методи. Для досягнення поставленої мети обрано амінокислотну послідовність TaWRKY2 (ID: ACD80357) з банку даних GenPept серверу NCBI з кількістю амінокислотних залишків 468 [2].

Серед методів гомологічного моделювання застосовували: Modeller [3], SWISS-MODEL[4], Phyre2 [5]. Серед методів з елементами *ab initio* використали D-I-TASSER [6]. Методи, засновані на машинному навчанні, включали AlphaFold2 [7, 8], RoseTTaFold [9], ESMFold [10] та OmegaFold [8].

Для оцінки якості отриманих моделей застосовували такі метрики, як Verify3D[11], ERRAT [12], Ramachandran Plot Analysis [12], G-factor [12], QMEAN [13], Z-Score [14], Molprobity Score [15], Clashscore [15].

Результати та обговорення. Серед підходів, які застосовували в дослідженні, найбільш якісні моделі з точки зору метрик (табл. 1) та біологічної функції отримано за допомогою методу AlphaFold2.

Зокрема з функціональної точки зору розташування двох доменів цинкових пальців у димері з послідовностями WRKYGQK, які розміщені ззовні димеру, дає їм змогу ефективно зв'язуватися з двома ділянками W-боксів ДНК (рис. 1).

В моделях, отриманих методами Phyre2, ESMFold, Modeller, присутні значні стеричні перешкоди для взаємодії з ДНК, метод RoseTTaFold змоделював некомпактну структуру, а метод SWISS-MODEL – неповну.

Таблиця 1. Порівняння метрик моделей, створених різними методами (позначка %-ль означає перцентиль, в який увійшла конкретна модель).

Метод\Метрика	Verify 3D, %	ERRAT	Ramachardran, %	G-factor	QMEAN	Z-Score	Molprobity Score/%-ль		Clashscore / %-ль	
SWISS-MODEL (неповна модель)	44.32	82.05	79.4	-0.38	0.35 ± 0.05	-3.23	1.82	84	0.59	99
Phyre2	69.66	3.91	57.6	-0.92	0.35 ± 0.05	0.05	3.65	6	221.26	0
AlphaFold (MMseqs2)	45.73	93.43	61,8	-0.75	0.38 ± 0.05	-3.72	2.19	65	0.72	99
AlphaFold (DeepMind)	48.93	85.23	62,8	-0.7	0.35 ± 0.05	-3.62	2.35	56	2.31	99
ESMFold	46.79	61.63	40,8	-2.4	0.39 ± 0.05	-4.13	3.64	6	24.15	22
RoseTTaFold	67.09	89.32	83,7	0.27	0.35 ± 0.05	-5.03	1.40	97	2.16	99
OmegaFold	46.37	32.72	44,7	-2.11	0.40 ± 0.05	-3.68	3.49	9	31.19	14
D-I-Tasser	50	65.42	51,4	-0.35	0.34 ± 0.05	-4.08	2.73	35	22.35	26
Modeller (TaWRKY19 як шаблон)	55.13	63.29	86,6	-0.31	0.40 ± 0.05	-2.98	2.21	64	16.15	45

Виконано 2 варіанта моделювання з використанням Google Colab та графічного процесора Google T4 GPU, включаючи AlphaFold2 з MMseqs2 пошуком (рис. 1) та AlphaFold2 від DeepMind (рис. 2) [8].

Обидві моделі мали високі (>0.9) значення внутрішньої метрики pLDDT (predicted Local Distance Difference Test) [7] для ділянок з цинковими пальцями і низькі (<0.5) для інших частин білку, що може бути спричинене наявністю неструктурованих або гнучких регіонів, відсутністю структурних аналогів у тренувальних даних, взаємодією з іншими молекулами та технічними обмеження методу AlphaFold.

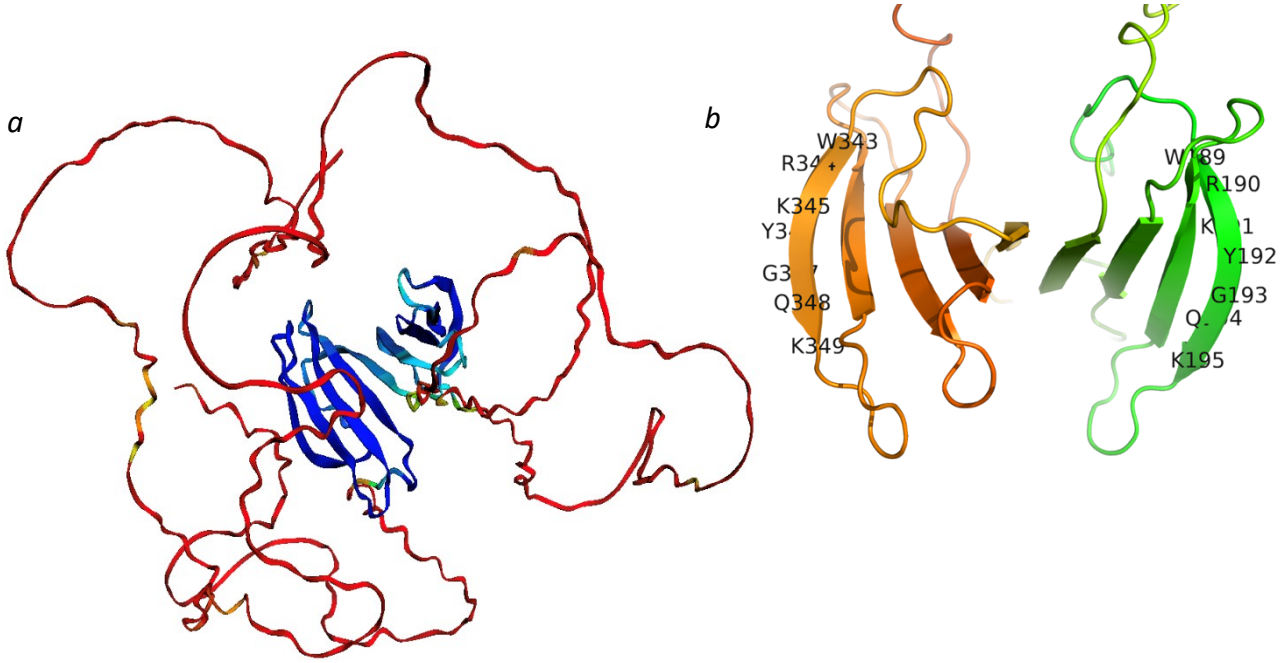

Рис. 1. Передбачена структура TaWRKY2 (AlphaFold, MMseqs2), забарвлено згідно з pLDDT (a); взаємне розташування послідовностей WRKYGQK в димері (b).

Аналіз показав, що модель AlphaFold (MMseqs2) мала кращі результати за метриками ERRAT, QMEAN, Z-Score, MolProbity Score та Clashscore (табл. 1), а AlphaFold (DeepMind) – за Verify3D, Ramachandran Plot, G-factor. Остання мала особливість – наявність додаткового β-ланцюга між двома WRKY-доменами, що робить структуру більш компактною і стабільною за рахунок великої кількості водневих зв'язків.

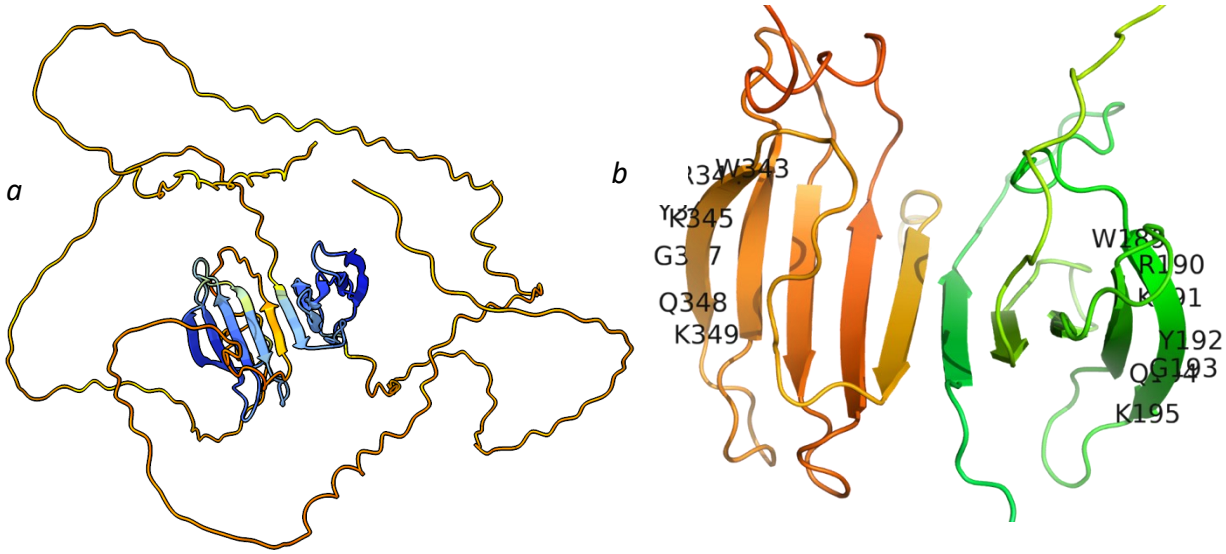

Рис. 2. Передбачена структура TaWRKY2 (AlphaFold, DeepMind), забарвлено згідно з pLDDT (a); взаємне розташування послідовностей WRKYGQK в димері (b).

Висновки. Після аналізу метрик моделей, отриманих різними методами, використання методу AlphaFold для передбачення структури транскрипційних факторів, зокрема TaWRKY2, дав найбільш достовірну просторову конфігурацію білка, оскільки ці моделі показали вищу якість та біологічну значущість.

Використання методів машинного навчання для передбачення експериментально не визначених структур дає змогу отримати глибше розуміння молекулярних механізмів у біологічних системах, зокрема стресостійкості.

Список використаної літератури:

1. Upadhyay, D., Budhlakoti, N., Kumari, J. *et al.* (2024). *In-silico* characterization of drought stress related *WRKY2* transcription factor in wheat crop (*Triticum aestivum* L.): study of its physico-chemical properties and structural dynamics. Genet Resour Crop Evol 71, 1481–1492.
2. https://www.ncbi.nlm.nih.gov/protein/ACD80357.1/
3. https://salilab.org/modeller/
4. https://swissmodel.expasy.org/
5. Muhammed Tilahun, Muhammed, Esin Aki-Yalcin. (2023). Up-to-Date Developments in Homology Modeling. Applied Computer-Aided Drug Design: Models and Methods, 116-135. Bentham Science.
6. Wei Zheng, Qiqige Wuyun, Yang Li, Quancheng Liu, Xiaogen Zhou, Yiheng Zhu, P. Lydia Freddolino, Yang Zhang. (2023). Integrating deep learning potentials with I-TASSER for single- and multi-domain protein structure prediction. Submitted.
7. Read, R. J., Baker, E. N., Bond, C. S., Garman, E. F., & van Raaij, M. J. (2023). AlphaFold and the future of structural biology. IUCrJ, *10*(Pt 4), 377–379.
8. Mirdita, M., Schütze, K., Moriwaki, Y. et al. (2022). ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682.
9. https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/
10. https://www.cgl.ucsf.edu/chimerax/docs/user/tools/esmfold.html
11. Marcin von Grotthuss, Jakub Pas, Lucjan Wyrwicz, Krzysztof Ginalski, Leszek Rychlewski (2003) Application of 3D-Jury, GRDB, and Verify3D in fold recognition. Proteins: Structure, Function, and Genetics, 53, 418-423.
12. Lahiri, T., Singh, K., Pal, M. K., & Verma, G. (2012). Protein structure validation using a semi-empirical method. Bioinformation, 8(20), 984–987.
13. Benkert, P., Künzli, M., & Schwede, T. (2009). QMEAN server for protein model quality estimation. Nucleic Acids Research, 37(suppl_2), W510–W514.
14. Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research, 35(suppl_2), W407–W410.
15. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall III, W. B., Snoeyink, J., Richardson, J. S., & Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research, 35(suppl_2), W375–W383.