

SOFTWARE PIPELINE FOR PREDICTING THE IMPACT OF MUTATIONS ON THE STABILITY OF PROTEIN SPATIAL STRUCTURES USING FREE ENERGY CHANGE ESTIMATION METHODS AND ARTIFICIAL INTELLIGENCE

A. S. Venzel, T. V. Ivanisenko, P. S. Demenkov, V. A. Ivanisenko

Institute of Cytology and Genetics, SB RAS,
630090, Novosibirsk, Russia

Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS,
630090, Novosibirsk, Russia
Novosibirsk State University,
630090, Novosibirsk, Russia

DOI: 10.24412/2073-0667-2024-4-6-16

EDN: EAMKIP

In this work, a computational pipeline was developed to predict the impact of mutations on the stability of protein structure. The pipeline employs a combined approach, utilizing state-of-the-art artificial intelligence methods for protein structure prediction and classical algorithms for free energy change estimation. The pipeline includes protein structure prediction using the ESM3 model, followed by calculation of free energy changes in mutant forms using pyRosetta. This approach allows overcoming the limitations of existing methods by combining the advantages of deep learning and the interpretability of energy calculations. The developed tool can find applications in structural bioinformatics, biotechnology, and medicine, especially given the limited number of experimentally determined protein structures.

Key words: protein structure prediction, protein structure stability, molecular modeling, ESM3.

References

1. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold // *Nature*. 2021. V. 596, № 7873. P. 583–589.
2. Abramson J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // *Nature*. 2024. P. 1–3.
3. Baek M. et al. Accurate prediction of protein structures and interactions using a three-track neural network // *Science*. 2021. V. 373, № 6557. P. 871–876.
4. Lin Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model // *Science*. 2023. V. 379, № 6637. P. 1123–1130.
5. Thomas P. J., Qu B. H., Pedersen P. L. Defective protein folding as a basis of human disease // *Trends in biochemical sciences*. 1995. V. 20, № 11. P. 456–459.
6. Kellogg E. H., Leaver-Fay A., Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability // *Proteins: Structure, Function, and Bioinformatics*. 2011. V. 79, № 3. P. 830–838.

The work is supported by a budget project of ICG SB RAS No FWNR-2022-0020.

7. Dehouck Y. et al. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality // BMC bioinformatics. 2011. V. 12. P. 1–12.
8. Schymkowitz J. et al. The FoldX web server: an online force field // Nucleic acids research. 2005. V. 33, № suppl_2. P. W382-W388.
9. Montanucci L. et al. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations // BMC bioinformatics. 2019. V. 20. P. 1–10.
10. Pires D. E. V., Ascher D. B., Blundell T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures // Bioinformatics. 2014. V. 30, № 3. P. 335–342.
11. Nikam R. et al. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years // Nucleic Acids Research. 2021. V. 49, № D1. P. D420-D424.
12. Xavier J. S. et al. ThermoMutDB: a thermodynamic database for missense mutations // Nucleic Acids Research. 2021. V. 49, № D1. P. D475-D479.
13. Stourac J. et al. FireProtDB: database of manually curated protein stability data // Nucleic Acids Research. 2021. V. 49, № D1. P. D319-D324.
14. Cao H. et al. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks // J. Chem. Inf. Model. 2019. V. 59, № 4. P. 1508–1514.
15. Umerenkov D. et al. PROSTATATA: a framework for protein stability assessment using transformers // Bioinformatics. 2023. V. 39, № 11. P. btad671.
16. Pak M. A. et al. Using AlphaFold to predict the impact of single mutations on protein stability and function // Plos one. 2023. V. 18, № 3. P. e0282689.
17. Mansoor S. et al. Zero-shot mutation effect prediction on protein stability and function using RoseTTAFold // Protein Science. 2023. V. 32, № 11. P. e4780.
18. Akdel M. et al. A structural biology community assessment of AlphaFold2 applications // Nature Structural & Molecular Biology. 2022. V. 29, № 11. P. 1056–1067.
19. Burley S.K. et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning // Nucleic Acids Research. 2023. V. 51, № D1. P. D488-D508.
20. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023 // Nucleic Acids Research. 2023. V. 51, № D1. P. D523-D531.
21. Hayes T. et al. Simulating 500 million years of evolution with a language model // bioRxiv. 2024. P. 2024.07.01.600583.
22. Frenz B. et al. Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy // Front. Bioeng. Biotechnol. 2020. V. 8.
23. Pancotti C. et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset // Briefings in Bioinformatics. 2022. V. 23. № 2. P. bbab555.
24. Chaudhury S., Lyskov S., Gray J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta // Bioinformatics. 2010. V. 26, № 5. P. 689–691.
25. Alford R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design // Journal of chemical theory and computation. 2017. V. 13, № 6. P. 3031–3048.
26. Zhang Y., Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score // Nucleic acids research. 2005. V. 33, № 7. P. 2302–2309.
27. Kunzmann P., Hamacher K. Biotite: a unifying open source computational biology framework in Python // BMC Bioinformatics. 2018. V. 19, № 1. P. 346.

ПРОГРАММНЫЙ КОНВЕЙЕР ПРЕДСКАЗАНИЯ ВЛИЯНИЯ МУТАЦИЙ НА СТАБИЛЬНОСТЬ ПРОСТРАНСТВЕННЫХ СТРУКТУР БЕЛКОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОЦЕНКИ ИЗМЕНЕНИЯ СВОБОДНОЙ ЭНЕРГИИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А. С. Вензель, Т. В. Иванисенко, П. С. Деменков, В. А. Иванисенко

Институт цитологии и генетики СО РАН,
630090, Новосибирск, Россия
Курчатовский геномный центр ИЦиГ СО РАН,
630090, Новосибирск, Россия
Новосибирский государственный университет,
630090, Новосибирск, Россия

УДК 575.112

DOI: 10.24412/2073-0667-2024-4-6-16

EDN: EAMKIP

В данной работе был разработан программный конвейер для предсказания влияния мутаций на стабильность пространственной структуры белка. В конвейере применяются комбинированный подход, использующий современные методы искусственного интеллекта для предсказания структуры белка, и классические алгоритмы оценки изменения свободной энергии. Конвейер включает в себя предсказание структуры белка с помощью модели ESM3 и последующий расчет изменения свободной энергии мутантных форм с помощью ruRosetta. Такой подход позволяет преодолеть ограничения существующих методов, объединяя преимущества глубокого обучения и интерпретируемость энергетических расчетов. Разработанный инструмент может найти применение в задачах структурной биоинформатики, биотехнологии и медицины, особенно в условиях ограниченного количества экспериментально определенных структур белков.

Ключевые слова: предсказание структуры белка, стабильность структуры белка, молекулярное моделирование, ESM3.

Введение. Активное развитие методов искусственного интеллекта для различных задач структурной биоинформатики сильно ускорило и облегчило процесс планирования экспериментов по получению белков с необходимыми функциями для задач биотехнологии и медицины. Наиболее заметный прогресс был достигнут в решении задачи предсказания структуры белка с использованием методов искусственного интеллекта, где такие методы как AlphaFold2 [1], AlphaFold3 [2], RoseTTAFold [3] и ESMFold [4] превзошли все существовавшие на тот момент классические методы, основанные на моделировании по гомологии.

Другой важной задачей структурной биоинформатики является предсказание эффекта влияния мутаций на стабильность пространственных структур белка. Стабильность

Работа поддержана бюджетным проектом ИЦиГ СО РАН № FWNR-2022-0020.

белка является важнейшим фактором, определяющим функцию белка, и тесно связана со свободной энергией свернутого состояния. Стабильная структура белка соответствует более низкому значению свободной энергии, в то время как дестабилизирующие мутации часто приводят к увеличению свободной энергии, потенциально приводя к неправильной укладке белка. В частности, нарушение функции белка в организме человека часто связано с прогрессированием генетических нарушений, нейродегенеративных заболеваний и рака [5].

Основной оценкой эффекта мутации на структуру белка является значение изменения свободной энергии структуры белка, обозначаемое $\Delta\Delta G$ и измеряемое в ккал/моль либо кДж/кг.

При оценке влияния мутаций на структуру, функцию и стабильность белка измененные формы белка сравниваются с белком дикого типа, который является не мутировавшей формой белка, являющейся нормальной формой белка для данного организма в естественных условиях. Это позволяет оценить, как мутации влияют на структуру, функцию и стабильность белка. Сами мутации в белках разделяются на два основных типа: прямые и обратные мутации. Прямые мутации — это изменения в последовательности аминокислот, при которых аминокислота дикого типа заменяется на другую. Обратные мутации возвращают мутированный белок к последовательности дикого типа.

Существующие методы предсказания влияния мутаций на стабильность белка можно разделить на две категории:

1) Методы, предсказывающие по структуре, на вход которым даются структура белка дикого типа и позиция мутации в последовательности белка, а также название аминокислоты мутанта. Самыми популярными методами являются Rosetta [6], PoP-MuSiC [7] и FoldX [8], основанные на расчетах энергетических потенциалов.

2) Методы, предсказывающие по последовательностям дикого типа и его мутантов. Примерами таких методов являются DDGun [9] и mCSM [10].

В то же время создание таких баз данных как ProThermDB [11], ThermoMutDB [12], FireProtDB [13] с экспериментальными значениями изменения стабильности белка у мутантов позволило создать множество методов глубокого обучения для предсказания изменения стабильности белков, используя как структуры, так и последовательности. Примерами таких моделей являются DeepDDG [14] и PROSTATATA [15].

Несмотря на высокую точность моделей глубокого обучения в прогнозировании изменений стабильности белков, они имеют два существенных недостатка: низкую интерпретируемость результатов и сильную зависимость от обучающей выборки данных. В отличие от них, методы, основанные на расчете энергетических потенциалов, позволяют разложить общую энергию на составляющие, соответствующие различным физико-химическим свойствам. Это дает возможность определить, какие именно типы молекулярных взаимодействий играют ключевую роль в стабилизации или дестабилизации белка при анализе результатов. В то же время методы, основанные на глубоком обучении, дают только финальное значение изменения свободной энергии $\Delta\Delta G$ без получения структуры мутанта.

Важно отметить, что ранее было показано, что упомянутые методы предсказания структур белков неэффективны для предсказания эффекта мутаций или структуры мутанта без дополнительного обучения модели [16, 17]. Однако было показано, что структуры белков, предсказанные AlphaFold2 и обладающие высокой точностью, могут быть использованы для предсказания эффекта мутаций [18].

Существует значительное неравенство между количеством экспериментально определенных структур белков в базе данных PDB (Protein Data Bank) (около 225 тысяч) [19] и числом известных белковых последовательностей в UniProt (порядка 250 миллионов) [20]. Это несоответствие подчеркивает острую необходимость в разработке надежных методов для предсказания структур белков и их мутантных форм, особенно для тех белков, чьи структуры еще не были экспериментально установлены.

В ходе этой работы был разработан программный конвейер, который предсказывает эффект мутации на свободную энергию пространственной структуры белка, предсказанной с помощью ESM3 [21]. В конвейере на языке программирования Python был реализован алгоритм оценки изменения свободной энергии у мутантных форм белка CartesianDDG из пакета макромолекулярного моделирования Rosetta [22].

1. Методы и материалы. Проверка конвейера и сравнение с другими методами проводились на независимой выборке данных s669 [23] для 669 прямых мутаций и соответствующих обратных мутаций у 94 структур белков. Для проверки были выбраны только случаи прямых мутаций. Пространственные структуры белков предсказаны с помощью предобученной генеративной мультимодальной языковой модели ESM3 с 7 миллиардами параметров: esm3_sm_open_v1 [21].

Для оценки качества предсказанных структур использовалась предсказанная ошибка *pLDDT* (predicted Local Distance Difference Test), которая является стандартной для методов предсказания структур и предоставляется самими методами предсказания вместе с предсказанными структурами. *pLDDT* оценивает от 0 до 100, насколько хорошо предсказание будет согласовываться с экспериментальной структурой на основе локального теста разницы расстояний между атомами C_α в структуре белка [1].

Минимизация структур выполнена с помощью протокола FastRelax из пакета макромолекулярного моделирования PyRosetta для языка программирования Python [24].

Внесение мутации и измерение изменения свободной энергии $\Delta\Delta G$ выполнены с помощью PyRosetta по алгоритму cartesian_ddg со скоринг-функцией ref2015 [25].

Структурное выравнивание пространственных структур белков выполнено с помощью TMAlign [26]. Анализ предсказанных структур выполнен с помощью библиотеки Biotite [27]. Конвейер проверялся на выборке данных s669 с экспериментальными значениями $\Delta\Delta G$ для 669 прямых мутаций у 94 структур белков из работы [23].

Метрики оценки других методов, также взяты из предыдущей работы. Оценка точности предсказания $\Delta\Delta G$ проводилась с помощью коэффициента корреляции Пирсона (r), квадратного корня средней квадратичной ошибки ($RMSE$) и средней абсолютной ошибки (MAE).

2. Результаты.

2.1. *Архитектура программного конвейера.* Был разработан программный конвейер (рис. 1), который начинает работу с инициализации пакета макромолекулярного моделирования PyRosetta с начальными параметрами: -ex1, -ex2, -flip_HNQ, -relax:cartesian, -relax:default_repeats 5, -nstruct 100, -optimization:default_max_cycles 100. Параметры -ex1 и -ex2 позволяют расширить пространство разрешенных конформационных изменений для углов χ_1 и χ_2 боковых радикалов, улучшая точность моделирования, -flip_HNQ позволяет поворачивать водороды амидной цепи для гистидина, аспарагина и глутамина, что обеспечивает правильное положение водородных связей, -relax: cartesian и -relax:default_repeats 5 повторяют 5 раз каждый процесс минимизации структуры в декартовых координатах, -nstruct 100 каждый процесс минимизации или внесения мутации

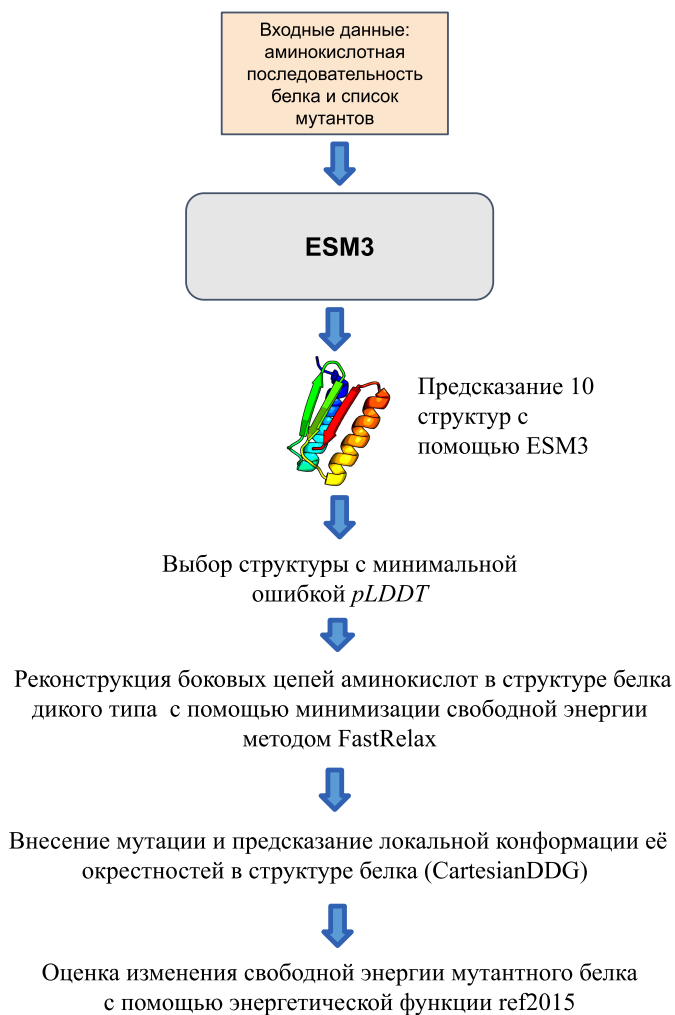


Рис. 1. Блок-схема работы программного конвейера

создает ансамбль из 100 структур с разными конформациями, откуда выбирается структура с минимальной энергией.

На вход конвейера подаются последовательность белка дикого типа и список мутаций, где каждая мутация записывается в следующем виде: A100B, где аминокислота A на 100 позиции в последовательности заменяется на аминокислоту B. Далее для последовательности дикого типа белка предсказывается 10 структур в ESM3 с разными заданными значениями параметра *seed* от 1 до 10. Из предсказанных структур выбирается структура с минимальным предсказанным средним значением ошибки *pLDDT*. Предсказанная структура имеет только основную цепь. Предсказанная структура дополняется боковой цепью и минимизируется с помощью алгоритма FastRelax со значением баростата 1 атм. и протоколом изменения конформации MonteCarlo, где в качестве оценочной функции энергии структуры использовалась функция ref2015, которая была параметризована так, чтобы воспроизводить термодинамически наблюдаемые и структурно-основанные свойства белков. Далее в минимизированную структуру вносятся мутации с помощью модифицированного алгоритма CartesianDDG (рис. 2). Изменение свободной энергии мутантных белков вычисляется по формуле:



Рис. 2. Блок-схема работы модифицированного алгоритма CartesianDDG в pyRosetta

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt},$$

где $\Delta\Delta G$ — изменение свободной энергии в REU (Rosetta Energy Units), ΔG_{mut} — свободная энергия мутанта, ΔG_{wt} — свободная энергия дикого типа. Далее $\Delta\Delta G$ были переведены в ккал/моль.

Кроме структур мутантов и значения $\Delta\Delta G$, в файл сохраняются значения энергетических терм из оценочной функции ref2015 для мутантных структур и для структуры дикого типа, которые подробно описаны в соответствующей работе [25].

2.2. *Результаты конвейера на s669.* s669 — набор данных, состоящий из 669 мутаций для 94 структур, а также экспериментальных значений $\Delta\Delta G$ для каждой прямой и обратной мутаций, при этом для каждого белка дикого типа присутствует экспериментальная структура в базе данных PDB, что дает возможность сравнить предсказанные структуры с экспериментальными. Были проверены только прямые мутации, так как ESM3, используемый в конвейере для предсказания структуры белка дикого типа, как и другие методы предсказания структур белков, был обучен на последовательностях дикого типа без учета мутантных форм, что может привести к неопределенному результату. Например, модели могут неправильно интерпретировать локальные эффекты мутации, распространяя их на всю структуру белка, когда в действительности эффект может быть более локализованным.

Каждая предсказанная структура дикого типа с помощью ESM3 была структурно выровнена на экспериментальную структуру из PDB, и вычислено значение $TM Score$ — относительное значение структурного сходства двух данных структур, вычисляемое TMAlign. Корреляция между средним значением ошибки предсказания $pLDDT$ и $TM Score$ составило $r = 0.82$ (рис. 3, а). При этом предсказанная структура (PDB ID: 2KJ3) с наименьшим $TM Score = 0.23$ имеет $pLDDT = 0.63$ (рис. 3, б), а структура (PDB ID: 1PRE) с наименьшим $pLDDT = 0.3$ имеет $TM Score = 0.24$ (рис. 3, в). При этом данная структура из рис. 3, б имеет допустимый показатель pLDDT, но при этом сама структура не обладает высокой точностью, что делает необходимым визуальную инспекцию каждой предсказанной структуры. Также среди предсказанных структур есть структуры с высо-

Таблица 1

Оценка методов предсказания изменения свободной энергии на s669, результаты представлены в виде коэффициента корреляции Пирсона r , среднеквадратичного отклонения (RMSE) и среднего оценочного отклонения (MAE). Результаты Rosetta взяты из Pancotti, 2022. pyRosetta — результаты реализованного алгоритма в pyRosetta алгоритм CartesianDDG из конвейера на структурах из PDB, ESM3 + pyRosetta — результаты конвейера с предсказанными структурами в ESM3

Метод	r	RMSE	MAE
Rosetta*	0.39	2.70	2.08
pyRosetta	0.36	2.90	2.20
ESM3+pyRosetta	0.10	5.77	3.47

Таблица 2

Результаты предсказания изменения свободной энергии программным конвейером, разделенные по категориям оцененной точности предсказанной структуры $pLDDT$

Категория точности предсказанной структуры дикого типа	Количество структур дикого типа	r	MAE	RMSE
Очень высокая ($pLDDT > 0.9$)	39	0.25	2.87	3.91
Достоверная ($0.9 > pLDDT > 0.7$)	36	0.07	2.99	5.31
Низкая ($0.7 > pLDDT > 0.5$)	11	-0.23	2.46	3.23
Очень низкая ($pLDDT < 0.5$)	42	-0.04	12.15	15.00

ким значением $TMScore$ и $pLDDT$ (рис. 3, г), в основном имеющие плотную глобулярную укладку.

При проверке части конвейера, выполняющей внесение мутации и подсчет изменения энергии на экспериментальных структурах из s669, были получены результаты, схожие с результатами Rosetta из Pancotti, 2022. При этом результаты, полученные на структурах, предсказанных ESM3, были значительно хуже (табл. 1). Причиной данного результата могут быть плохо предсказанные структуры.

Предсказанные структуры можно разделить на 4 категории по качеству предсказания в зависимости от средней ошибки $pLDDT$: на структуры, предсказанные с очень высокой оценочной точностью ($pLDDT > 0.9$), достоверной ($0.9 > pLDDT > 0.7$), низкой ($0.7 > pLDDT > 0.5$), очень низкой ($pLDDT < 0.5$). В результате разделения предсказанных структур белков дикого типа по категориям качества предсказанной структуры (табл. 2) корреляция между экспериментальным $\Delta\Delta G$ и предсказанным $\Delta\Delta G$ мутантов для структур с очень высокой точностью ($pLDDT > 0.9$) сильно увеличивается ($r = 0.25$), но становится хуже при $pLDDT < 0.9$.

Заключение. В этой работе был разработан программный конвейер предсказания влияния мутаций на стабильность предсказанных структур белков с использованием методов оценки изменения свободной энергии. На вход конвейеру подается последовательность белка дикого типа и список мутаций. В ходе работы конвейера сначала предсказывается структура белка с помощью мультимодальной генеративной модели ESM3, в которую далее вносится мутация и рассчитываются изменения в свободной энергии у мутанта. В результате работы конвейера пользователь получает структуры мутантов и значение из-

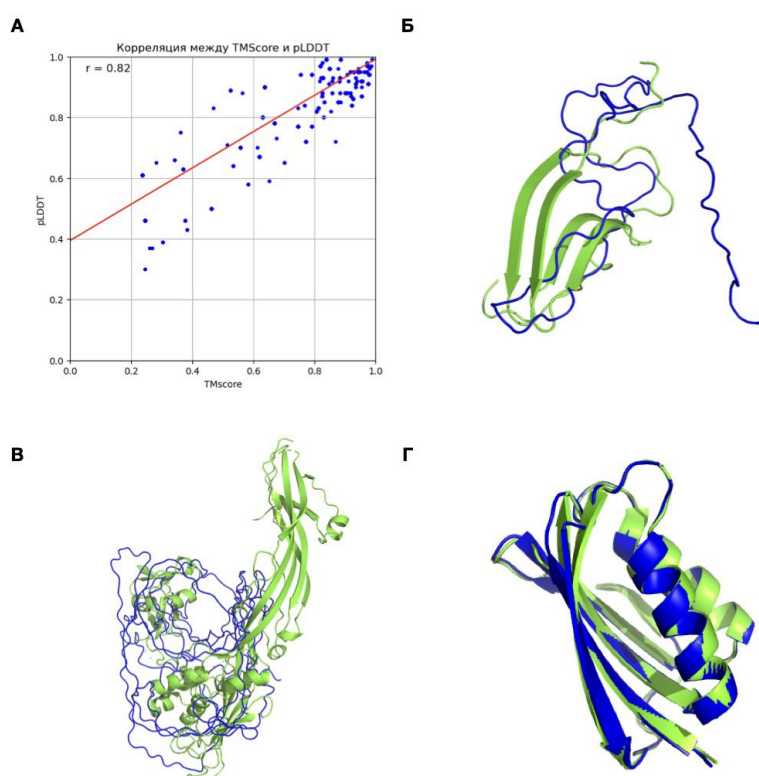


Рис. 3. Результаты предсказания структур в ESM3. (А) график корреляции между $pLDDT$ и $TMScore$, (Б) сравнение предсказанной структуры (синий) 2KJ3 с экспериментальным (зеленый): $TMScore = 0.23$, $pLDDT = 0.63$, (В) сравнение предсказанной структуры (синий) 1PRE с экспериментальным (зеленый): $pLDDT = 0.3$, $TMScore = 0.24$, (Г) сравнение предсказанной структуры (синий) 2BJD с экспериментальным (зеленый): $TMScore = 0.99$, $pLDDT = 0.99$

менения свободной энергии, а также изменения отдельных энергетических терм оценочной функции.

Анализ точности результатов разработанного конвейера проводился на наборе данных s669. Часть конвейера, ответственная за расчет влияния мутации на свободную энергию структуры, тестировалась отдельно на экспериментально полученных структурах белков PDB из набора данных s699 и показала схожий результат с ранее полученным результатом Rancotti, 2022 для данного алгоритма. Предсказание влияния мутаций на структуру показало сильную зависимость от качества предсказанной структуры, оцениваемое мерой $pLDDT$. Для достижения минимальной ошибки предсказания изменения свободной энергии мутации должны вноситься в структуры с очень высоким значением $pLDDT > 0.9$, при значениях $pLDDT < 0.9$ предсказанные структуры нужно оценить визуально либо каким-то другим методом. При этом присутствует сильная корреляция между средним $pLDDT$ предсказанных структур и значением $TMScore$, характеризующим структурное сходство предсказанной структуры с экспериментальной, из чего можно предположить, что при отсутствии экспериментальной структуры белка предсказанная структура с высоким $pLDDT$ может быть использована в качестве замены.

Список литературы

1. Jumper J. и др. Highly accurate protein structure prediction with AlphaFold // *Nature*. 2021. Т. 596. № 7873. С. 583–589.
2. Abramson J. и др. Accurate structure prediction of biomolecular interactions with AlphaFold 3 // *Nature*. 2024. С. 1–3.
3. Baek M. и др. Accurate prediction of protein structures and interactions using a three-track neural network // *Science*. 2021. Т. 373. № 6557. С. 871–876.
4. Lin Z. и др. Evolutionary-scale prediction of atomic-level protein structure with a language model // *Science*. 2023. Т. 379. № 6637. С. 1123–1130.
5. Thomas P. J., Qu B. H., Pedersen P. L. Defective protein folding as a basis of human disease // *Trends in biochemical sciences*. 1995. Т. 20. № 11. С. 456–459.
6. Kellogg E. H., Leaver-Fay A., Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability // *Proteins: Structure, Function, and Bioinformatics*. 2011. Т. 79. № 3. С. 830–838.
7. Dehouck Y. и др. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality // *BMC bioinformatics*. 2011. Т. 12. С. 1–12.
8. Schymkowitz J. и др. The FoldX web server: an online force field // *Nucleic acids research*. 2005. Т. 33. № suppl_2. С. W382–W388.
9. Montanucci L. и др. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations // *BMC bioinformatics*. 2019. Т. 20. С. 1–10.
10. Pires D. E. V., Ascher D. B., Blundell T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures // *Bioinformatics*. 2014. Т. 30. № 3. С. 335–342.
11. Nikam R. и др. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years // *Nucleic Acids Research*. 2021. Т. 49, № D1. С. D420–D424.
12. Xavier J.S. et al. ThermoMutDB: a thermodynamic database for missense mutations // *Nucleic Acids Research*. 2021. Т. 49, № D1. С. D475–D479.
13. Stourac J. и др. FireProtDB: database of manually curated protein stability data // *Nucleic Acids Research*. 2021. Т. 49, № D1. С. D319–D324.
14. Cao H. и др. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks // *J. Chem. Inf. Model. American Chemical Society*, 2019. Т. 59, № 4. С. 1508–1514.
15. Umerenkov D. и др. PROSTAT: a framework for protein stability assessment using transformers // *Bioinformatics*. 2023. Т. 39, № 11. С. btad671.
16. Pak M. A. и др. Using AlphaFold to predict the impact of single mutations on protein stability and function // *Plos one*. 2023. Т. 18. № 3. С. e0282689.
17. Mansoor S. и др. Zero-shot mutation effect prediction on protein stability and function using RoseTTAFold // *Protein Science*. 2023. Т. 32, № 11. С. e4780.
18. Akdel M. и др. A structural biology community assessment of AlphaFold2 applications // *Nature Structural & Molecular Biology*. 2022. Т. 29. № 11. С. 1056–1067.
19. Burley S.K. и др. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning // *Nucleic Acids Research*. 2023. Т. 51, № D1. С. D488–D508.
20. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023 // *Nucleic Acids Research*. 2023. Т. 51, № D1. С. D523–D531.
21. Hayes T. и др. Simulating 500 million years of evolution with a language model // *bioRxiv*. 2024. С. 2024.07. 01.600583.
22. Frenz B. и др. Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy // *Front. Bioeng. Biotechnol. Frontiers*, 2020. Т. 8.

23. Pancotti C. et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset // *Briefings in Bioinformatics*. 2022. Т. 23. № 2. С. bbab555.

24. Chaudhury S., Lyskov S., Gray J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta // *Bioinformatics*. 2010. Т. 26, № 5. С. 689–691.

25. Alford R. F. и др. The Rosetta all-atom energy function for macromolecular modeling and design // *Journal of chemical theory and computation*. 2017. Т. 13. № 6. С. 3031–3048.

26. Zhang Y., Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score // *Nucleic acids research*. 2005. Т. 33. № 7. С. 2302–2309.

27. Kunzmann P., Hamacher K. Biotite: a unifying open source computational biology framework in Python // *BMC Bioinformatics*. 2018. Т. 19, № 1. С. 346.



Вензель Артур Сергеевич — аспирант, младший научный сотрудник Института цитологии и генетики СО РАН. Области научных интересов: биоинформатика, структурная биология, ИИ в биологии. E-mail: venzel@bionet.nsc.ru.

Artur Sergeevich Venzel — PhD student, junior researcher at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Areas of scientific interest: bioinformatics, structural biology, AI in biology. E-mail: venzel@bionet.nsc.ru.

тересов: биоинформатика, генные сети, системная биология, большие геномные данные, ИИ в биологии, text-mining. E-mail: demps@bionet.nsc.ru.



Pavel Sergeevich

Demenkov — PhD in Computer Science, researcher at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Areas of scientific interest: bioinformatics, gene networks, systems biology, big genomic data, AI in biology text mining. E-mail: demps@bionet.nsc.ru.



Иванисенко Тимофей Владимирович — научный сотрудник Института цитологии и генетики СО РАН. Области научных интересов: биоинформатика, генные сети, системная биология, большие геномные данные, ИИ в биологии, text-mining. E-mail: itv@bionet.nsc.ru.

Timovey Vladimirovich Ivanisenko — researcher at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Areas of scientific interest: bioinformatics, gene networks, systems biology, big genomic data, AI in biology text mining. E-mail: itv@bionet.nsc.ru.

Деменков Павел Сергеевич — канд. техн. наук, научный сотрудник Института цитологии и генетики СО РАН. Окончил НГУ в 2005 году по специальности «Прикладная математика и информатика». Защитил кандидатскую диссертацию в 2008 году. Области научных ин-

Иванисенко Владимир Александрович — канд. биол. наук, доцент, заведующий лабораторией компьютерной протеомики и лабораторией искусственного интеллекта и больших геномных данных ИЦиГ СО РАН. Области научных интересов: биоинформатика, генные сети, структурная биология, системная биология, большие геномные данные, ИИ в биологии, text-mining. E-mail: salix@bionet.nsc.ru.

Vladimir Alexandrovich Ivanisenko — PhD in Biology, associate professor, the head of the Laboratory of Computational Proteomics and the Laboratory of Artificial Intelligence and Big Genomic Data at the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. Areas of scientific interest: bioinformatics, gene networks, structural biology, systems biology, big genomic data, AI in biology text mining. E-mail: salix@bionet.nsc.ru.