ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И СВОРАЧИВАНИЕ БЕЛКОВ: ОТ ФУНДАМЕНТАЛЬНОЙ ПРОБЛЕМЫ БИОФИЗИКИ К БИОМОЛЕКУЛЯРНОМУ КОНСТРУИРОВАНИЮ

Б. Ишанхаджаев

Филиал Астраханского государственного технического университета в Ташкентской области

Обзорная статья посвящена анализу прорыва в области молекулярной биологии и биоинформатики, вызванного применением технологий искусственного интеллекта для решения проблемы сворачивания белков. В работе рассмотрен исторический контекст проблемы - от парадокса Левинталя и гипотезы Анфинсена до международных соревнований CASP. Подробно анализируется эволюция вычислительных методов: архитектура системы AlphaFold 2 на базе модуля Evoformer и механизмов внимания, а также генеративная модель AlphaFold 3, применяющая диффузионные алгоритмы для предсказания структуры комплексных биомолекулярных систем (белок–белок, белок–ДНК/РНК, белок–лиганд).

Ключевые слова: искусственный интеллект, AlphaFold, сворачивание белков, третичная структура, диффузионные модели, Evoformer, биоинформатика, рациональный дизайн лекарств.

This review article analyzes the breakthrough in molecular biology and bioinformatics driven by the application of artificial intelligence technologies to solve the protein folding problem. The paper examines the historical context of the problem, ranging from Anfinsen's dogma and Levinthal's paradox to the CASP experiment. The evolution of computational methods is discussed in detail, focusing on the AlphaFold 2 architecture based on the Evoformer module and attention mechanisms, as well as the generative AlphaFold 3 model, which employs diffusion algorithms to predict the structures of complex biomolecular assemblies (protein-protein, protein-DNA/RNA, protein-ligand).

Keywords: artificial intelligence, AlphaFold, protein folding, tertiary structure, diffusion models, Evoformer, bioinformatics, rational drug design.

Введение и исторический контекст

Кристиан Анфинсен, получая Нобелевскую премию по химии за исследования рибонуклеазы, высказал гипотезу, ставшую краеугольным камнем молекулярной биологии: трехмерная нативная структура белка полностью определяется его аминокислотной последовательностью. Однако путь от одномерной цепи аминокислот к трехмерной конформации оказался крайне сложным.

Американский физик Сайрус Левинталь сформулировал знаменитый парадокс: если бы полипептидная цепь из 100 аминокислот искала свою правильную пространственную форму путем случайного перебора всех возможных конформаций, на это потребовалось бы около 10^{27} лет — период, превышающий возраст Вселенной. В живой клетке этот процесс занимает миллисекунды.

Первичная структура (аминокислотная цепь)

|

▼ [Проблема фолдинга / Парадокс Левинталя]

Трехмерная третичная структура (активный белок)

На протяжении полувека определение структуры белков оставалось прерогативой **экспериментальных методов:**

рентгеноструктурный анализ (РСА): требует получения высокочистых кристаллов белка, что не всегда возможно (особенно для мембранных белков);

ядерный магнитный резонанс (ЯМР): ограничен молекулярной массой исследуемого объекта;

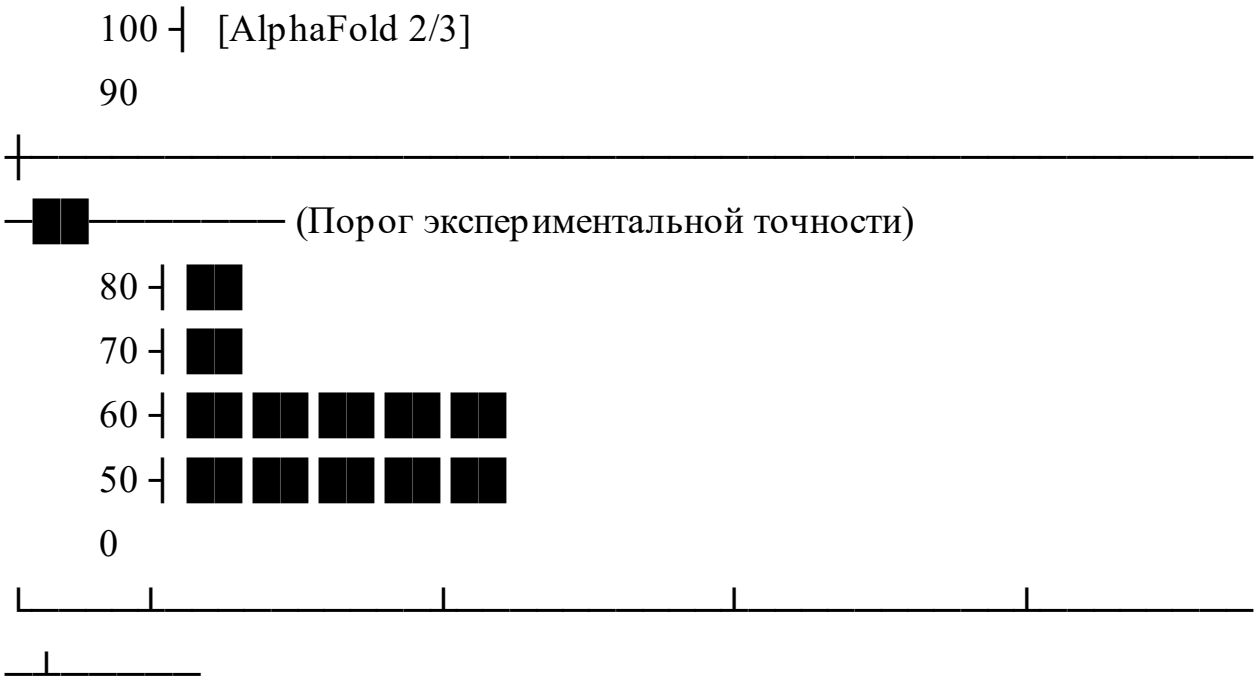
криоэлектронная микроскопия (Крио-ЭМ): требует дорогостоящего оборудования и длительной подготовки образцов.

За 50 лет экспериментальных исследований ученые смогли расшифровать лишь небольшую часть изтенных белков в природе. Возникла остросюжетная потребность в вычислительных методах, способных предсказывать структуру белков с точностью, сопоставимой с экспериментальной.

Компьютерная революция: алгоритмы и CASP

Для объективной оценки прогресса в биоинформатике в 1994 году был основан эксперимент CASP (Critical Assessment of Structure Prediction). Раз в два года исследовательские группы получают последовательности аминокислот белков, чьи структуры уже определены экспериментально, но еще не опубликованы в базе данных Protein Data Bank (PDB).

Ключевой метрикой оценки точности является тест GDT (Global Distance Test), измеряемый от 0 до 100. Значение GDT выше 90 означает, что предсказанная модель практически идентична экспериментально полученной структуре. Точность моделей на CASP (GDT Score, 2006–2024):



До 2018 года показатели лучших программ едва достигали 60–65 баллов GDT. Переломный момент наступил на CASP13 (2018 г.), когда команда лаборатории DeepMind (Google) представила алгоритм AlphaFold 1, а затем на CASP14 (2020 г.) - AlphaFold 2, преодолевший отметку в 90 баллов GDT и решивший проблему предсказания структуры отдельной белковой цепи.

Архитектура AlphaFold 2 и AlphaFold 3.

AlphaFold 2. В основе AlphaFold 2 лежит архитектура нейросети Evoformer и модуль Structure Module:

1. Множественное выравнивание последовательностей (MSA): нейросеть ищет эволюционно родственные белки в генетических базах данных. Парные мутации аминокислот в процессе эволюции указывают на то, что эти остатки находятся близко друг к другу в трехмерном пространстве.

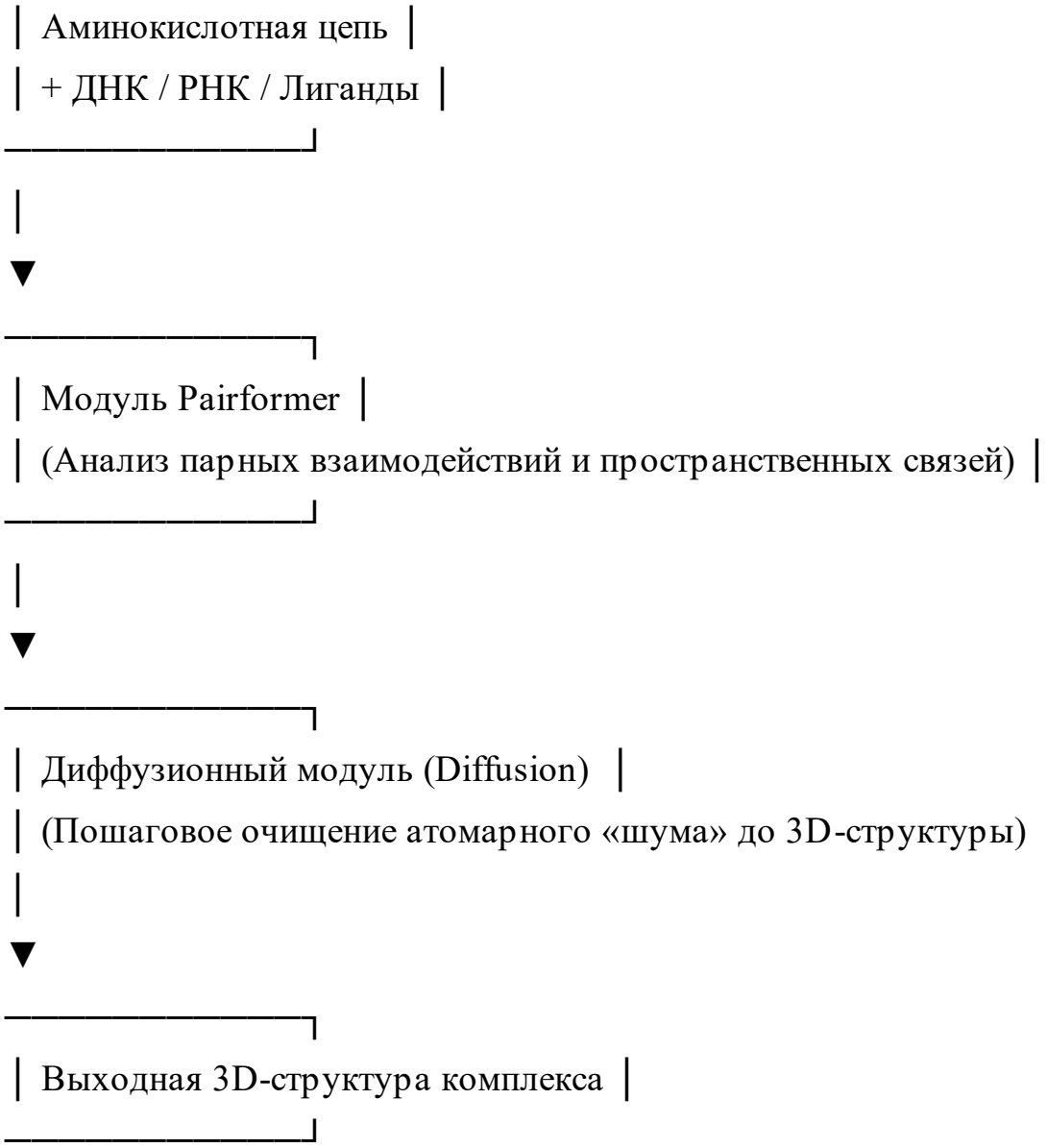
2. Evoformer: анализирует эволюционные связи и пространственные контакты аминокислот одновременно через механизмы внимания (Attention Mechanism).

3. Structure Module: преобразует полученные представления в трехмерные координаты атомов полипептидного скелета и боковых цепей.

AlphaFold 3 (2024 год). В 2024 году команда DeepMind представила AlphaFold 3 - генеративную модель нового поколения. Главные отличия:

переход к диффузионным моделям: Модуль Evoformer уступил место более простому Pairformer, а финальные 3D-координаты генерируются диффузионной сетью (похожей на DALL-E и Midjourney, но работающей с координатами атомов);

универсальность: AlphaFold 3 прогнозирует структуры не только одиночных белков, но и сложных комплексов: белок–белок, белок–ДНК/РНК, белок–лиганд (лекарство) и посттрансляционные модификации.



Практическое применение в науке и медицине. Развитие систем ИИ в структурной биологии принесло прикладные результаты в различных областях биомедицины:

Область	Применение
Рациональный дизайн лекарств	Моделирование взаимодействия молекул-кандидатов в лекарственные препараты с активными центрами белков-мишеней (вирусных ферментов).
Синтетическая биология	Проектирование новых ферментов, разрушающих пластик (ПЭТ), или эффективных биокатализаторов для промышленности.
Медицинская	Анализ точечных мутаций: оценка того, как

генетика	однонуклеотидный полиморфизм меняет конформацию белка и приводит к заболеванию.
Разработка вакцин	Моделирование структуры поверхностных антигенов вирусов и бактерий для создания стабильных иммуногенов.

Открытие доступа к базе данных AlphaFold Protein Structure Database сделало доступными структуры более 200 миллионов белков - практически всех известных науке видов.

Ограничения и перспективы. Несмотря на фундаментальные успехи, современный ИИ в биофизике сталкивается со сбоями и ограничениями:

1. Динамика и конформационные переходы: белки - это не жесткие статуэтки, а динамические молекулярные машины. ИИ предсказывает преимущественно одно, наиболее стабильное энергетическое состояние.

2. Внутренне неструктурированные белки (IDP): значительная часть белков эукариот не имеет постоянной структуры в свободном состоянии и сворачивается только при взаимодействии с партнером.

3. Зависимость от обучающей выборки: ИИ опирается на данные PDB. Предсказание принципиально новых, не похожих ни на что искусственных полимеров может вызывать трудности.

4. Перспективы: будущее связано с объединением ИИ с молекулярной динамикой (MD) для моделирования белковых процессов в реальном времени, а также с интеграцией экспериментальных данных cryo-EM высокого разрешения.

Список литературы

1. Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).
2. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
3.

4. Baek, M., DiMaio, F., Anishchenko, I. et al. Accurate prediction of protein structures and complexes using RoseTTAFold. Science 373 (6557), 871–876 (2021).
5. Dill, K. A., MacCallum, J. L. The Protein-Folding Problem, 50 Years On. Science 338 (6110), 1042–1046 (2012).