ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ / ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI: https://doi.org/10.60797/COMP.2025.5.1

ПРИМЕНЕНИЕ МЕТОДОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА В ОБЛАСТИ ЯЗЫКОВЫХ МОДЕЛЕЙ БЕЛКОВ: ТЕКУЩИЕ И БУДУЩИЕ ТЕНДЕНЦИИ

Научная статья

Шарнин М.М.[1, *], Сомин Н.В.[2]
[1] ORCID : 0000-0003-0450-5156;
[2] ORCID : 0000-0002-8683-4617;
[1, 2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Российская Федерация

* Корреспондирующий автор (mc[at]keywen.com)

Аннотация

В работе представлен прогнозный библиометрический анализ трендовых тем в коллекции PubMed в области языковых моделей белков. Анализ выполнен с использованием коллекции научных статей PubMed, из которой были отобраны 187 статей, имеющие в заголовках слова «protein» (белок) и «language» (язык). Выявлен значительный рост (в 54 раза за 6 лет) ежегодно публикуемых подобных статей. Рассчитан и представлен рейтинг релевантных ключевых слов в отобранных статьях. Среди релевантных ключевых слов выявлены трендовые ключевые слова прогнозируемым долгосрочным ростом трендов. Представлена семантическая карта трендовых ключевых слов, содержащая информацию о новизне и долгосрочности трендов. В результате визуального анализа семантической карты выявлены четыре трендовые темы:

1) обработка естественного языка (natural language processing);
2) базы данных и языковые модели белков (databases and protein language models);
3) глубокое обучение (deep learning);
4) белковая инженерия (protein engineering).

Дано сравнение выявленных трендов с опубликованными в научной литературе.

Ключевые слова: обработка естественного языка, языковые модели белков, библиометрический анализ, долгосрочный прогноз трендов, визуальная аналитика, трендовые темы, коллекция PUBMED.

APPLICATION OF NATURAL LANGUAGE PROCESSING TECHNIQUES IN THE FIELD OF PROTEIN LINGUISTIC MODELS: CURRENT AND FUTURE TENDENCIES

Research article

Sharnin M.M.[1, *], Somin N.V.[2]
[1] ORCID : 0000-0003-0450-5156;
[2] ORCID : 0000-0002-8683-4617;
[1, 2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russian Federation

* Corresponding author (mc[at]keywen.com)

Abstract

The work presents a predictive bibliometric analysis of trending topics in the PubMed collection in the field of linguistic models of proteins. The analysis was performed using the PubMed collection of scientific articles, from which 187 articles with the words "protein" and "language" in their titles were selected. A significant increase (54-fold over 6 years) in the number of such articles published annually was found. The ranking of relevant keywords in the selected articles was calculated and presented. Among the relevant keywords, trending keywords with predicted long-term trend growth are identified. A semantic map of trending keywords containing information about the novelty and long-term trend growth is presented. As a result of visual analysis of the semantic map, four trending topics are identified:

1) natural language processing;
2) databases and protein language models;
3) deep learning;
4) protein engineering.

The detected trends are compared with those published in the scientific literature.

Keywords: natural language processing, protein language models, bibliometric analysis, long-term trend forecasting, visual analytics, trending topics, PUBMED collection.

Введение

В последние годы наблюдается быстрое развитие задач обработки естественного языка (natural language processing, NLP). В большинстве случаев, оно связано с идеей представления слова/термина в виде вектора, который вычисляется на основе контекста и потому несет информацию о его семантике. Получаемые из текста с помощью нейронных сетей векторные представления называются встраиваниями (embeddings), т.к. они в совокупности образуют семантическое векторное пространство, в которое эти вектора встроены. Векторное представление оказалось очень удобным для

вычисления похожести слов и фрагментов текста, в результате чего появились многочисленные разработки NLP, в том числе следующие.

Рекурентные нейронные сети (recurrent neural networks, RNN) [1] — популярный вид нейронных сетей, используемых в обработке естественного языка. Такие сети появились еще в 80-х годах. Особенностью RNN является переменное число внутренних слоев сети – по длине цепочки слов. Поэтому RNN сети обычно применяют в задачах, где вход распадается на произвольное число токенов, например, для распознавания рукописных текстов или перевода на другие языки.

Позже появилась архитектура Трансформера (Transformer) [2], уже не требующая строгой последовательности в обработке входа. Трансформер состоит из двух частей кодировщика (encoder) и декодировщика (decoder). Такая архитектура, как правило, использовалась в последующих разработках. Но уже в первых Трансформерах был реализован алгоритм внимания (attention), повышающий веса наиболее важных при переводе слов.

В настоящее время NLP-нейросети активно используются в системах искусственного интеллекта, разработанных крупными фирмами. Так, нейронная сеть BERT (Bidirectional Encoder Representations from Transformers) [3], разработанная компанией Google, была встроена в знаменитый текстовый поисковик этой фирмы, что позволило лучше понимать поисковые запросы. Для этого сеть была обучена на огромном количестве текстов (2500 млн слов) и может понимать 104 языка. Другая сеть GPT-4 (Generative Pre-trained Transformer 4) [4], созданная фирмой OpenAI, четвёртая в серии GPT. Она была выпущена 14 марта 2023 года и сразу стала доступна для многочисленных пользователей известной ИИ-среды ChatGPT.

Методы и системы, применяемые в NLP, неожиданно нашли широкое применение в задачах анализа и синтеза белков. Дело в том, что каждую аминокислоту, из которых состоят белки, можно рассматривать как токен, а всю белковую цепочку – как текст. Совсем недавно языковые модели белков (protein language models, PLM) с неконтролируемым обучением были дополнительно исследованы для извлечения признаков из большого объема последовательностей белков. Интересно, что такие предварительно обученные большие PLM выдают признаки белков, содержащие богатые структурные и функциональные свойства белков, и было доказано, что они очень эффективны во многих задачах исследования белков, таких как прогнозирование вторичной структуры и мутационных эффектов.

Языковые модели белков стремительно развиваются. За последние годы количество ежегодно публикуемых статей в области языковых моделей белков выросло в несколько десятков раз. Однако до сих пор выявление трендов в данной предметной области выполняется на основе экспертных оценок, причем не приводятся данные о долгосрочности и точности прогноза для выявленных будущих трендов. Настоящая работа призвана исправить этот недостаток. В ней авторы стремились к достижению трех основных целей исследования:

1) из имеющихся в литературе работ по языковым моделям белков выявить трендовые ключевые слова, имеющие наиболее долговременные растущие тренды по количеству статей и цитирований;

2) выявить и визуализировать трендовые темы (тенденции) из близких по семантике трендовых ключевых слов;

3) с помощью предлагаемого в работе метода выявить наиболее перспективные трендовые темы и получить для них оценки долгосрочности и точности прогноза.

Обзор связанных работ

В научной литературе имеется ряд работ с анализом трендов в области языковых моделей белков. В данном обзоре особое внимание будем уделять текущим и будущим трендам в этой области, а также долгосрочности будущих трендов, если такая информация имеется. Вначале рассмотрим более ранние работы.

Ofer и др., 2021 [5] отмечали, что важным источником вдохновения для биоинформатики становятся методы обработки естественного языка (NLP), такие как глубокое обучение, нейронные языковые модели, генеративные модели, самоконтролируемое обучение и модели, основанные на внимании. Обильные и высококачественные данные играют важную роль в исследовании белков и предсказании структуры белков. Наличие стандартизированных, объективных критериев для сравнения методов имеет решающее значение для сосредоточения усилий на наиболее перспективных методах и идеях.

Bepler и др., 2021 [6] исследовали язык белков с точки зрения, эволюции, структуры и функций. Авторы отметили, что биологические последовательности не являются естественным языком, и необходимо разработать новые языковые модели, отражающие уникальные для белков свойства и фундаментальную природу биологических последовательностей. В будущем глубокие генеративные модели белков смогут моделировать белки, образующиеся в результате эволюционных процессов. Увеличение размера модели, вычислительной мощности и размера набора данных улучшит производительность и качество языковых моделей белков. Этому также будет способствовать дополнение моделей специфическими свойствами белков, такими как структура и функция.

Ferruz и др., 2022 [7] отмечали, что последовательности белков по своей сути похожи на естественные языки. Внедрение предварительно обученных моделей Трансформеров позволило генерировать текст с возможностями, подобными человеческим. Авторы ожидают, что специализированные Трансформеры будут доминировать в генерации индивидуальных последовательностей белков в ближайшем будущем. Авторы считают, что использование генеративных текстовых моделей для создания новых белков является многообещающей областью, и позволит осуществить разработку белков с настраиваемыми свойствами, что является давней целью в биохимии.

Qiu и др., 2023 [8] сделали обзор работ в области белковой инженерии с использованием искусственного интеллекта, включая топологический анализ данных и глубокие языковые модели белков. Авторы отмечают, что такие модели могут извлекать критическую эволюционную информацию из крупномасштабных баз данных белковых последовательностей. Накопленные базы данных белков и модели машинного обучения, основанные на обработке естественного языка (NLP), значительно ускорили белковую инженерию. В будущем различные виды машинного и глубокого обучения будут иметь большие перспективы в белковой инженерии.

Vu и др., 2023 [9] предложили вдохновленную лингвистикой дорожную карту для создания биологически надежных языковых моделей белков. Авторы отмечали, что глубокие нейронные сети на основе языковых моделей все чаще применяются к крупномасштабным данным последовательностей белков для прогнозирования функций белков. Включение лингвистических идей в языковые модели белков позволяет разрабатывать интерпретируемые модели следующего поколения с потенциалом раскрытия биологических механизмов, лежащих в основе отношений типа последовательность-функция.

Savojardo и др., 2023 [10] обсуждали поиск функциональных мотивов в последовательностях белков с помощью глубокого обучения и моделей естественного языка. Они отмечали, что языковые модели белков (PLMs) являются мощными подходами для извлечения информации об эволюции. Авторы указывали на необходимость внешних сравнительных критериев (бенчмаркинга) и достаточного объема обучающих/тестовых данных.

Huang и др., 2023 [11] проанализировали текущий прогресс, проблемы и будущие перспективы языковых моделей для представления и дизайна белков. Авторы считают, что контролируемый дизайн белков произведет революцию в разработке лекарств. В числе будущих проблем проектирования белков с помощью моделей глубокого обучения авторы отмечают необходимость в высококачественных данных и необходимость учитывать возможность динамического изменения структуры белков. Авторы полагают, что языковые модели помогут понять процессы сворачивания белков и их биологические функции.

Heinzinger и др., 2023 [12] описали двуязычную языковую модель последовательности и структуры белков и их перспективы. Авторы отмечают двойственную природу белков, которые действуют и существуют как трехмерные (3D) машины и развиваются в виде линейных цепочек одномерных (1D) последовательностей. Авторы отмечают необходимость одновременно использовать обе модальности, комбинируя 1D-последовательности с 3D-структурой в одной общей модели. В будущем большие языковые модели (LLM) смогут интегрировать любые существующие сегодня знания об определенном белке, включая информацию из научных статей и онтологии генов (Gene Ontology).

Zhang и др., 2024 [13] сделали обзор применений больших научных лингвистических моделей в биологической и химической областях. Авторы отмечают, что большие языковые модели (LLM) стали преобразующей силой в улучшении понимания естественного языка. Среди перспективных направлений будущих исследований авторы отмечают: создание крупномасштабных, высококачественных и кросс-модальных наборов обучающих данных, разработку метрик и сравнительных критериев, включение в LLM внешних знаний, и 3D-стереовременной информации.

Из приведенного обзора можно сделать следующие выводы. В рассмотренных статьях прогнозы по будущим трендам определялись на основе экспертных мнений/оценок. Часто в качестве будущих трендов выбирались наиболее активные текущие тренды. При этом не приводились данные о долгосрочности прогнозируемых будущих трендов и точности этих прогнозов. Данная статья предоставляет эту недостающую информацию.

Постановка задачи

Область исследования данной статьи – это публикации, касающиеся языковых моделей белков. Настоящая статья посвящена прогнозу и визуализации трендовых тем в исследуемой области.

Если говорить более точно, то ставится следующая задача: дать прогноз долгосрочности роста трендов с известной точностью на 3 и более лет вперед для трендовых тем в области исследования на базе коллекции PubMed, а также предложить средства визуализации прогноза трендовых тем.

3.1. Исходный текстовый материал

Для решения поставленной задачи была проанализирована известная текстовая коллекция по медицинской тематике PubMed, которая по состоянию на начало 2024 года содержала более 36 миллионов статей (36 417 711) по медицине, биологии и связанным наукам. Из этой коллекции было выделено 187 статей, содержащие в заголовках слова «protein» (белок) и «language» (язык). Данные этих 187 статей мы называем в дальнейшем Локальной коллекцией.

Таким образом в рамках данной работы мы работаем с двумя коллекциями: с коллекцией PubMed, содержащей 36,417,711 записей/статей, и с Локальной коллекцией, содержащей 187 статей. Локальная коллекция является частью коллекции PubMed, причем частью, наиболее сильно связанной с исследуемой областью. Свойства ключевого слова из Локальной коллекции (частота, вероятность, тренды, контекст и т.д.) в дальнейшем называются локальными, а соответствующие свойства этого же ключевого слова в рамках коллекции PubMed называются глобальными. Как локальные, так и глобальные свойства необходимы для выявления характерных/релевантных и трендовых ключевых слов.

3.2. Актуальность

Актуальность поставленной задачи является несомненной, т.к. количество статей в исследуемой области лавинообразно растет. По данным коллекции PubMed был рассчитан следующий график роста количества статей в сфере языковых моделей белков по годам (см. рис. 1).

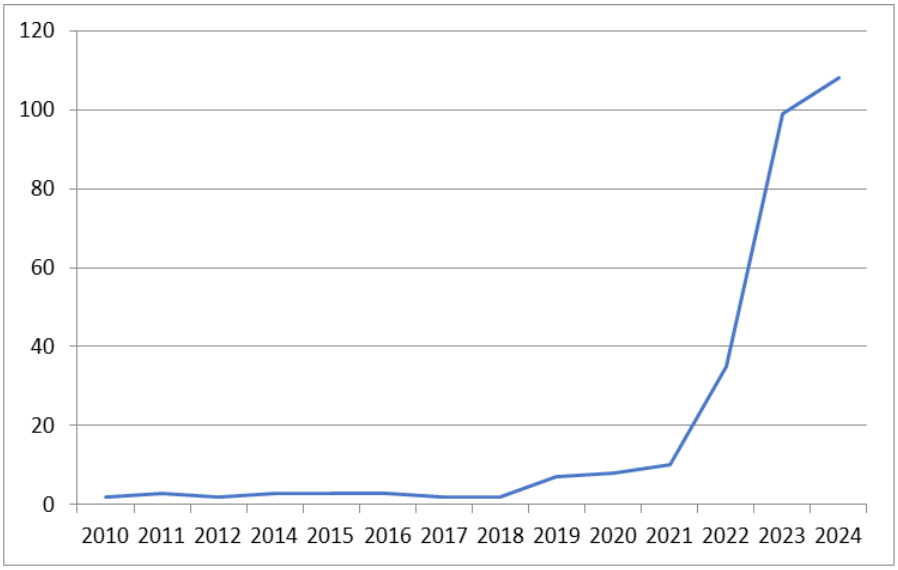

Рисунок 1 - График количества научных статей в сфере языковых моделей белков за различные годы по данным PubMed на август 2024 года

DOI: https://doi.org/10.60797/COMP.2025.5.1.1

Из рисунка 1 видно, что наблюдается резкий рост количества ежегодно публикуемых научных статей, индексируемых в PubMed, в области языковых моделей белков. Точнее, с 2010 по август 2024 количество статей в этой области выросло в 54 раза и в настоящее время их более 300.

Методы исследования

Рассмотрим описываемые в статье этапы анализа, основные методы исследования и используемые программные средства.

4.1. Этапы анализа

Для выявления и визуализации трендовых тем используется библиометрический анализ, в котором находят применение ряд математических и статистических методов к изучению библиографической коллекции PubMed. Библиометрический анализ исследуемой области содержит следующие этапы:

1) определение поисковых запросов, с помощью которых можно найти статьи из PubMed, относящиеся к исследуемой области,

2) построение Локальной коллекции, в которую включены заголовки статей из PubMed, удовлетворяющих поисковым запросам, и анализ динамики их количества за последние несколько лет,

3) выявление в полученном материале релевантных, а затем и трендовых ключевых слов с упорядочиванием их по степени перспективности,

4) построение семантической карты и выявление трендовых тем (тенденций) с помощью визуальной аналитики,

5) обзор наиболее перспективных трендовых тем и сравнение их с темами, выявленными другими авторами.

4.2. Основные методы исследования и применяемые программные средства

Трендовые ключевые слова определяются с учетом долгосрочных прогнозов их трендов в PubMed, рассчитанных с помощью пакета машинного обучения CatBoost [14]. Авторская методика долгосрочного прогноза трендов описана в работах [15], [16]. Точность прогноза на 3 года вперед составляет более 60%. При этом 59% трендовых слов, выявленных в 2020 году, остались трендовыми в 2023 году.

Результаты прогноза трендов визуализируются с помощью нейросети Word2Vec и алгоритма t-SNE на семантической карте. С помощью визуальной аналитики выявляются трендовые темы, входящие в кластеры трендовых слов на семантической карте. Трендовая тема состоит из одного или нескольких близких по смыслу трендовых ключевых слов. Выявленные будущие тренды сравниваются с уже опубликованными в научной литературе.

Результаты библиометрического анализа

Результаты библиометрического анализа включают в себя рейтинг релевантных ключевых слов, рейтинг трендовых ключевых слов и семантическую карту.

5.1 Рейтинг релевантных ключевых слов

Характерными/релевантными в исследуемой области являются те ключевые слова, у которых локальная вероятность вхождения в заголовки статей выше чем аналогичная глобальная вероятность. Локальная вероятность определяется по Локальной коллекции, а глобальная – по коллекции PubMed.

Верхние позиции в рейтинге релевантных ключевых слов занимают следующие ключевые слова: protein language models (языковые модели белков), protein language (язык белков), language model (языковая модель), language models (языковые модели), language (язык), pre-trained (предварительно обученный), pretrained (предварительно обученный), pretrained language model (предварительно обученная языковая модель), natural language (естественный язык), embeddings (встраивания), biological language (биологический язык), languages (языки), language modeling (моделирование языка), model embeddings (встраивания модели), embeddings protein (встраивания белковой), protein

sequences (последовательности белков), language life (язык живых), markup language (язык разметки), markup (разметка), language processing (обработка языка), structures families (семейства структур), language specification (спецификация языка), protein markup (разметка белкового), specification protein (спецификации белковой), extracting protein-protein (извлечение белок-белковых), dynamical language (динамический язык), pretrained protein (предварительно обученная белковая), transformer protein (трансформер-модель белков), modelling language (язык моделирования), deep protein (глубокая белковая), language text (текст языка), pre-trained language (предварительно обученная языковая), prediction protein (предсказания белковых), protein (белок), site prediction (прогнозирование сайта), query language (язык запросов), variant effects (эффекты вариантов), interaction extraction (извлечение взаимодействий), learning protein (обучение белковых) и т.д. Из этих ключевых слов удалены предлоги.

Данный рейтинг представляет наиболее релевантные ключевые слова из 242 ключевых слов без предлогов, упоминающихся в Локальной коллекции три или более раз.

5.2 Рейтинг трендовых ключевых слов

Трендовые ключевые слова были отобраны среди релевантных ключевых слов с учетом их локальных параметров в Локальной коллекции и глобальных параметров в коллекции PubMed. Следует заметить, что при составлении рейтинга анализируются показатели/вероятности для слов, пар слов, и троек из соседних слов без предлогов.

С помощью алгоритма машинного обучения CatBoost по методике, описанной в предыдущем разделе, были рассчитаны прогнозы долгосрочности роста трендов для релевантных слов и выявлены трендовые ключевых слова. Отобранные ключевые слова были упорядочены по совокупности показателей и в результате получился рейтинг трендовых ключевых слов. Рейтинг составлялся с учетом следующих показателей: положение в рейтинге характерных слов, данные прогноза, частота в коллекции PubMed и в Локальной коллекции, а также по тенденциям роста в Локальной коллекции.

Верхние позиции в рейтинге трендовых ключевых слов занимают следующие термины: pre-trained (предварительно обученный), deep learning (глубокое обучение), ensemble (ансамбль), graph, (граф), structure prediction (прогнозирование структуры), multiple sequence (множественная последовательность), prediction protein (предсказание белковой), ensemble learning (ансамблевое обучение), protein structure (структура белка), integrating (интеграция), embeddings (встраивания), neural networks (нейронные сети), protein-protein (белок-белок), protein-protein interactions (взаимодействия белок-белок), machine learning (машинное обучение), protein-protein interaction (взаимодействие белок-белок), robust (надежный), fast accurate (быстрый точный), protein engineering (инженерия белка), learning protein (обучение белковой), networks protein (нейросети белковых), protein sequences (последовательности белка), variant effects (вариантные эффекты), site prediction (предсказание сайта), models protein (модели белковых), natural language (естественный язык), language modeling (моделирование языка), language processing (обработка языка), protein design (проектирование белка), modelling (моделирование), discovering (открытие), database (база данных), structured (структурированный), semantic (семантический), online (онлайн), speech-language (язык речи), programming (программирование), specification (спецификация), interaction extraction (извлечение взаимодействия) и т.д. У первых 20 слов из этого списка (pre-trained от до learning protein) прогноз роста более 3 лет (т.е. после 2026 года).

5.3 Построение семантической карты

Для целей визуального анализа с помощью нейросети Word2Vec была рассчитана мера семантического подобия (semantic similarity) трендовых ключевых слов в коллекции PubMed. По этим данным с помощью алгоритма t-SNE (см. рис. 2) построена семантическая карта. Чем выше мера подобия – тем меньше расстояние между ключевыми словами на семантической карте.

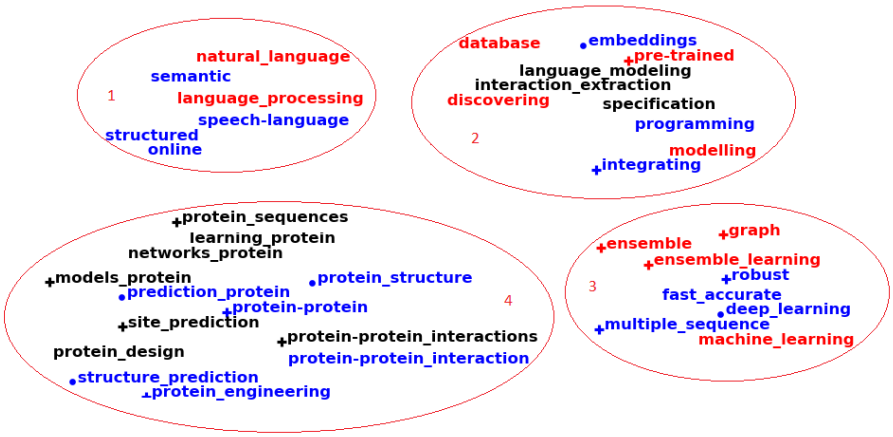

Рисунок 2 - Семантическая карта трендовых ключевых слов в области языковых моделей белков

DOI: https://doi.org/10.60797/COMP.2025.5.1.2

На рис. 2 красным и синим цветами представлены трендовые термины PubMed. Красным цветом выделены наиболее перспективные термины, имеющие самые долгосрочные тренды в коллекции PubMed, синим – средние тренды, черным – минимальные тренды. Плюсом (+) выделены ключевые слова, имеющие самые новые/свежие тренды, а точкой – средние по новизне тренды.

На рис. 2 также видны четыре кластера, которые соответствуют четырем трендовым темам и которые содержат близкие по семантике трендовые термины. На основе визуального анализа выявлены четыре трендовые темы:

1) обработка естественного языка (natural language processing);
2) базы данных и языковые модели белков (databases and protein language models);
3) глубокое обучение (deep learning);
4) белковая инженерия (protein engineering).

Визуальный анализ семантической карты помогает определить трендовые темы, оценить их динамику и увидеть картину в целом, включая перспективные направления из наиболее перспективных тем. Например, из рис.2 видно, что темы/кластеры «глубокое обучение» и «базы данных и языковые модели белков» представляют перспективные направления, т.к. они содержат больше перспективных ключевых слов, выделенных красным цветом.

Сравнение выявленных трендов с опубликованными

Рассмотрим четыре трендовые темы в сфере языковых моделей белков, выявленные в процессе прогнозного библиометрического анализа, и сравним выявленные будущие тренды с уже опубликованными в научной литературе. Также дадим краткие пояснения для некоторых новых терминов в области искусственного интеллекта и биотехнологий.

К первой теме «обработка естественного языка» (natural language processing) относятся статьи, в которых обсуждается применение методов обработки естественного языка, включая большие языковые модели (LLM), к языковым моделям белков. Эта тема имеет следующие ключевые слова: естественный язык (natural language), обработка языка (language processing), обработка естественного языка (natural language processing), NLP, семантический (semantic). Обработка естественного языка упоминается в качестве трендовой темы в семи работах [5], [7], [10], [13], описанных в Разделе 2. Методы обработки естественного языка оказались полезными для различных целей, в том числе для: разработки белков с настраиваемыми свойствами [7], извлечения эволюционной информации [6], [8], поиска функциональных мотивов в белковых последовательностях [10], разработки интерпретируемых моделей [9], обнаружения сходства белков [17], а также для прогнозирования взаимодействий белков [18].

Вторая трендовая тема «базы данных и языковые модели белков» посвящена базам данных о белках и построению на их основе языковых моделей белков (PLM), включая предварительно обученные языковые модели (pre-trained language models). Эта тема имеет ключевые слова: база данных (database), моделирование языка (language modeling), предварительно обученный (pre-trained), предварительно обученные языковые модели (pre-trained language models), встраивание (embeddings), интеграция (integrating), обнаружение (discovering). Базы данных и языковые модели белков упоминаются в качестве трендовых тем в восьми работах [5], [8], [11], [13], описанных в Разделе 2.

Рассмотрим некоторые дополнительные статьи из этой темы. Pratyush и др., 2023 [19] использовали интеграцию контролируемого встраивания (embeddings) слов и встраивания из предварительно обученной модели языка белков для прогнозирования участков S-нитрозилирования белков, что важно для понимания механизмов клеточной сигнализации как у животных, так и у растений. Haselbeck и др., 2023 [20] использовали встраивания (embeddings) из языковых моделей белков для предсказания термостабильности белков, что важно для ферментной инженерии и белково-гибридной оптоэлектроники.

Третья трендовая тема «глубокое обучение» (deep learning) содержит статьи, в которых описываются различные нейросетевые методы построения языковых моделей белков. Эта тема имеет трендовые ключевые слова: глубокое обучение (deep learning), машинное обучение (machine learning), ансамблевое обучение (ensemble learning), граф, графовый (graph). Надо сказать, что графы очень удобны для представления пространственной/геометрической структуры белков. Потенциальным решением проблемы анализа сложной структуры графов является представлений графов в низкомерном пространстве с помощью методов встраивания графов (graph embedding). Модели глубокого обучения на графах (например, графовые нейронные сети) продемонстрировали превосходную производительность в различных задачах [21].

Глубокое обучение (deep learning) и его разновидности упоминается в качестве трендовых тем в работах [5], [7], [9], [11]. Рассмотрим некоторые дополнительные статьи из этой темы. Wang и др., [22] использовали графовую сверточную нейросеть (graph convolution network) и языковую модель белка для прогнозирования эффектов мутаций на растворимость белка, снижение которой (растворимости) может приводить к заболеваниям. Hou и др., 2023 [23] использовали объяснимое ансамблевое глубокое обучение (ensemble deep learning) для изучения языка белков в местах связывания белок-белок, что важно при функциональном анализе белков.

Наконец, четвертая трендовая тема «белковая инженерия» (protein engineering) содержит статьи, связанные с разработкой новых белков (protein design), прогнозированим структуры белков (prediction of protein structure), прогнозированим взаимодействий белков (prediction of protein interactions). Эта тема имеет ключевые слова: белковая инженерия (protein engineering), дизайн белков (protein design), структура белка (protein structure), последовательности белка (protein sequences), прогнозирование белковой (prediction of protein), прогнозирование структуры (structure prediction), белок-белковые взаимодействия (protein-protein interactions).

Белковая инженерия и ее составляющие упоминаются в качестве трендовой темы в пяти работах [5], [6], [8], [11], описанных в Разделе 2. Рассмотрим некоторые дополнительные статьи из этой темы. Lei и др., 2023 [18] использовали методы обработки естественного языка для предсказания взаимодействий типа белок-белок у растений-патогенов, что важно для понимания патогенной инфекции и иммунитета растений. Frisby и др., 2023 [24] использовали глубокую языковую модель белков на базе Трансформера для идентификации перспективных белковых последовательностей, которые с наибольшей вероятностью обладают целевыми, желаемыми свойствами. Qiu и др., 2023 [8] утверждают что, белковая инженерия может произвести революцию в различных областях, таких как разработка антител, разработка лекарственных препаратов, продовольственная безопасность и экология.

Из приведенного выше сравнения видно, что большинство выявленных будущих трендов подтверждаются уже опубликованными научными работами, описанными в Разделе 2. В данном сравнении также приведены дополнительные тренды, являющиеся уточнениями трендовых тем, например, графы (graph), графовая сверточная нейросеть (graph convolution network) и встраивания графов (graph embedding).

Заключение

Методы обработки естественного языка (NLP), нашли широкое применение в области биотехнологий, в задачах анализа и синтеза белков. В последние годы наблюдается резкий рост количества работ по языковым моделям белков. Причем особенно стремительно развиваются такие методы и темы/направления, как использование глубокого обучения, графовые нейронные сети, а также и предварительно обученные языковые модели.

В данной работе в результате прогнозного библиометрического анализа выявлены и рассмотрены следующие четыре трендовые темы в области языковых моделей белков:

1) обработка естественного языка (natural language processing);
2) базы данных и языковые модели белков (databases and protein language models);
3) глубокое обучение (deep learning);
4) белковая инженерия (protein engineering).

Методы NLP оказались полезными для построения нейросетевых языковых моделей белков. Развитие языковых моделей белков, в свою очередь, в будущем окажет позитивное влияние на методы NLP, т.к. специфика биотехнологий и возможность там экспериментальных проверок является плодотворной средой для совершенствования языковых моделей, что будет полезно как для самих биотехнологий, так и для обработки естественного языка. Данное явление может быть названо трансфером технологий между биотехнологиями и лингвистикой.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network / A. Sherstinsky // Physica D: Nonlinear Phenomena. — 2020. — № 404. — P. 132306.
2. Lin T. A survey of transformers / T. Lin, Y. Wang, X. Liu [et al.] // AI open. — 2022. — № 3. — P. 111–132.
3. Kenton J.D.M.W.C. Bert: Pre-training of deep bidirectional transformers for language understanding / J.D.M.W.C. Kenton, L.K. Toutanova // Proceedings of naacL-HLT. — 2019. — Vol. 1. — P. 2.
4. Liu X. GPT understands, too / X. Liu, Y. Zheng, Z. Du [et al.] // AI Open. — 2023.
5. Ofer D. The language of proteins: NLP, machine learning & protein sequences / D. Ofer, N. Brandes, M. Linial // Computational and Structural Biotechnology Journal. — 2021. — № 19. — P. 1750–1758.
6. Bepler T. Learning the protein language: Evolution, structure, and function / T. Bepler, B. Berger // Cell systems. — 2021. — № 12 (6). — P. 654–669.
7. Ferruz N. Controllable protein design with language models / N. Ferruz, B. Höcker // Nature Machine Intelligence. — 2022. — № 4 (6). — P. 521–532.
8. Qiu Y. Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models / Y. Qiu, G.W. Wei // Briefings in bioinformatics. — 2023. — № 24 (5). — P. bbad289.
9. Vu M.H. Linguistically inspired roadmap for building biologically reliable protein language models / M.H. Vu, R. Akbar , P.A. Robert [et al.] // Nature Machine Intelligence. — 2023. — № 5 (5). — P. 485–496.
10. Savojardo C. Finding functional motifs in protein sequences with deep learning and natural language models / C. Savojardo, P.L. Martelli, R. Casadio // Current Opinion in Structural Biology. — 2023. — № 81. — P. 102641.
11. Huang T. Current progress, challenges, and future perspectives of language models for protein representation and protein design / T. Huang, Y. Li // The Innovation. — 2023. — № 4 (4).
12. Heinzinger M. Bilingual language model for protein sequence and structure / M. Heinzinger, K. Weissenow, J.G. Sanchez [et al.] // bioRxiv. — 2023.
13. Zhang Q. Scientific large language models: A survey on biological & chemical domains / Q. Zhang, K. Ding, T. Lyv [et al.] // arXiv preprint. — 2024. — arXiv: 2401.14656.
14. Prokhorenkova L. CatBoost: unbiased boosting with categorical features / L. Prokhorenkova, G. Gusev, A. Vorobev [et al.] // Advances in neural information processing systems. — 2018. — Vol. 31.
15. Charnine M. Research trending topic prediction as cognitive enhancement / M. Charnine, A. Klokov, L. Kochiev [et al.] // 2021 international conference on cyberworlds (CW). IEEE. — 2021. — P. 217–220.
16. Charnine M. Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis / M. Charnine, A. Tishchenko, L. Kochiev // Proceedings of the 31th International Conference on Computer Graphics and Vision. — 2021. — Vol. 2. — P. 1028–1037.
17. Sarkar I.N. Discovering protein similarity using natural language processing / I.N. Sarkar, T.C. Rindflesch // Proceedings of the AMIA Symposium. American Medical Informatics Association. — 2002. — P. 677.

18. Lei C. AraPathogen2. 0: An Improved Prediction of Plant–Pathogen Protein–Protein Interactions Empowered by the Natural Language Processing Technique / C. Lei, K. Zhou, J. Zheng [et al.] // Journal of Proteome Research. — 2023. — № 23 (1). — P. 494–499.

19. Pratyush P. pLMSNOSite: an ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model / P. Pratyush, S. Pokharel, H. Saigo [et al.] // BMC bioinformatics. — 2023. — № 24 (1). — P. 41.

20. Haselbeck F. Superior protein thermophilicity prediction with protein language model embeddings / F. Haselbeck, M. John, Y. Zhang [et al.] // NAR Genomics and Bioinformatics. — 2023. — № 5 (4). — P. lqad087.

21. Zhang S. Graph convolutional networks: a comprehensive review / S. Zhang, H. Tong, J. Xu [et al.] // Computational Social Networks. — 2019. — № 6 (1). — P. 1–23.

22. Wang J. Predicting the effects of mutations on protein solubility using graph convolution network and protein language model representation / J. Wang, S. Chen, Q. Yuan [et al.] // Journal of Computational Chemistry. — 2024. — № 45 (8). — P. 436–445.

23. Hou Z. Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning / Z. Hou, Y. Yang, Z. Ma // Communications Biology. — 2023. — № 6 (1). — P. 73.

24. Frisby T.S. Identifying promising sequences for protein engineering using a deep transformer protein language model / T.S. Frisby, C.J. Langmead // Proteins: Structure, Function, and Bioinformatics. — 2023. — № 91 (11). — P. 1471–1486.