Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2026;30(2):293-298

doi 10.18699/vjgb-26-33

Alembic: от разрозненных биологических данных к структурированным ресурсам

И.В. Бездворных#, К.И. Юдыцкий#, Н.А. Черкасов, А.А. Самсонова, А.А. Канапин

Институт трансляционной биомедицины, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
a.kanapin@gmail.com

Аннотация. Развитие технологий высокопроизводительного секвенирования и методов анализа больших данных создает устойчивую потребность в повторном анализе накопленной в открытых репозиториях гетерогенной информации. Серьезной проблемой при этом остается преобладание свободного текстового описания биологических экспериментов, что затрудняет продуктивный поиск, систематизацию и дальнейшее использование соответствующих наборов данных. Прогресс в области искусственного интеллекта, особенно в развитии методов обработки естественного языка (natural language processing, NLP), обуславливает новые методологические возможности для эффективного решения этой задачи. Интегрированная система баз данных Entrez, поддерживаемая Национальным центром биотехнологической информации США (NCBI), предоставляет развитый и надежный доступ как к исходным данным секвенирования, так и к сопутствующей метаинформации, включающей детальное описание параметров экспериментов, через программный интерфейс (application programming interface, API). Это позволяет идентифицировать и загружать данные секвенирования и соответствующие им метаданные с описаниями экспериментов, используя поиск по ключевым словам и различным терминам, таким, например, как имена генов, в репозиториях; преобразовывать и систематизировать текстовые описания с применением современных NLP-методов и обеспечивать исследователям структурированную информацию для интеграции в локальные базы данных и форматированный перечень ссылок для загрузки исходных данных. Программный пакет Alembic предлагает комплексное решение для поиска и загрузки данных, автоматизируя все указанные этапы. Платформа использует клиент-серверную архитектуру и предназначена для локальной установки. Для анализа биомедицинских текстов, сопровождающих данные секвенирования, в Alembic интегрированы современные алгоритмы искусственного интеллекта на основе архитектуры трансформеров. В частности, используется имеющаяся в открытом доступе платформа AIONER, обученная на данных репозитория PubMed с помощью модели PubMedBERT. Такой подход обеспечивает эффективное распознавание именованных сущностей (named entity recognition, NER) биомедицинского характера (гены, заболевания и др.), предоставляя пользователю структурированные результаты поиска по ключевым словам. Формируемый пакетом список дает возможность исследователю анализировать результаты, отбирать наиболее релевантные наборы данных и получать всю необходимую информацию (включая исходные данные) для создания локального репозитория, ориентированного на конкретную исследовательскую задачу. В отличие от имеющихся аналогов, Alembic является универсальным решением для интеграции данных из репозиториев открытого доступа и работы с разнородными типами данных секвенирования.

Ключевые слова: обработка естественных языков; анализ биомедицинских текстов; семантическая аннотация; гармонизация данных; интеграция омиксных даных

Для цитирования: Бездворных И.В., Юдыцкий К.И., Черкасов Н.А., Самсонова А.А., Канапин А.А. Alembic: от разрозненных биологических данных к структурированным ресурсам. *Вавиловский журнал генетики и селекции*. 2026;30(2):293-298. doi 10.18699/vjgb-26-33

Финансирование. Работа поддержана грантом Российского научного фонда 23-14-00134.

Alembic: a framework for converting disparate biological data into structured resources

I.V. Bezdvornykh#, K.I. Yuditskiy#, N.A. Cherkasov, A.A. Samsonova, A.A. Kanapin

Institute for Translational Biomedicine, Saint Petersburg State University, St. Petersburg, Russia
a.kanapin@gmail.com

Abstract. The imperative to re-analyze existing public sequencing data is central to modern biology, driven by new hypotheses and advanced analytical methods. However, this effort is critically hampered by the profound heterogeneity of repository data, particularly the non-standardized, free-text descriptions of biological experiments. This lack of structural and semantic homogeneity prevents systematic search, integration, and comparative analysis, effectively locking away the full potential of accumulated datasets. Advances in Natural Language Processing (NLP) offer a pivotal pathway to overcome this bottleneck by transforming unstructured text into computable, homogeneous information.

Авторы внесли равный вклад в исследование.

The integrated Entrez database system, maintained by the National Center for Biotechnology Information (NCBI), provides sophisticated programmatic access via an API to primary sequencing data and its associated metadata, including detailed experimental descriptions. This interface enables researchers to identify and retrieve relevant data through keyword searches, including those based on gene names, and to apply modern NLP techniques to transform textual metadata into structured information. The output is formatted data ready for integration into local databases, accompanied by a systematic list of links for downloading primary files. The Alembic software package offers a comprehensive and automated solution for the entire workflow. Designed as a locally deployable client-server system, Alembic incorporates state-of-the-art transformer-based AI algorithms for analyzing the biomedical text that accompanies sequencing data. Its core utilizes the openly available AIONER platform, which is built upon the PubMedBERT model trained on the PubMed repository, to ensure efficient and accurate recognition of biomedical named entities (e. g., genes, diseases). This provides users with structured and meaningful keyword search results. By delivering a curated list of datasets, Alembic streamlines the path from search to analysis. Researchers can efficiently identify high-value targets and obtain a complete package of metadata and primary data to construct a tailored local repository. This positions Alembic as a universal solution that overcomes the fragmented approach of existing tools, offering an integrated workflow for diverse public sequencing data.

Key words: natural language processing; biomedical text mining; semantic annotation; data harmonization; omics data integration

For citation: Bezdvornykh I.V., Yuditskiy K.I., Cherkasov N.A., Samsonova A.A., Kanapin A.A. Alembic: a framework for converting disparate biological data into structured resources. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2026;30(2):293-298. doi 10.18699/vjgb-26-33

Введение

Данные секвенирования различных модальностей (таких как WGSeq, RNASeq, BSSeq) и сопутствующие публикации накапливаются со все возрастающей скоростью, что напрямую связано с развитием технологий высокопроизводительного секвенирования. Обработка таких массивов информации требует создания специализированных программных систем для автоматического эффективного извлечения биологической информации. Перспективным направлением в этой области стало применение алгоритмов обработки естественного языка (natural language processing, NLP). Современные NLP-методы автоматизируют анализ научных текстов, существенно повышая эффективность работы с большими данными. Одним из примеров применения NLP в биологии выступает система MetaMap (Aronson, Lang, 2010), реализующая распознавание именованных сущностей (named entity recognition, NER) для идентификации концептов метатезауруса UMLS (Unified Medical Language System) в текстах. Дальнейшее развитие этого направления воплотилось в пакете BIONER для распознавания биомедицинских сущностей (Wang et al., 2019).

Появление нейросетевых архитектур и техники векторных представлений слов (word embeddings) существенно повысило точность распознавания биомедицинских сущностей. Значительным шагом стало внедрение контекстнозависимых моделей типа ELMo (embeddings from language model), формирующих представления слов с учетом их окружения. Современный этап развития связан с доминированием трансформерных архитектур, прежде всего BERT (Devlin et al., 2019), кардинально изменивших методы решения NLP-задач. Дальнейшая адаптация этих моделей к биомедицинской предметной области привела к созданию домен-специфичных решений (BioBERT, PubMedBERT и др.) (Lee et al., 2020) и специализированных инструментов преобразования данных, таких как scispaCy (Neumann et al., 2019).

Наиболее комплексной системой биомедицинских баз данных остается Entrez, разрабатываемая Национальным центром биотехнологической информации США (NCBI). Платформа объединяет 38 репозиториев, включая PubMed, PMC, базы нуклеотидных и белковых последовательностей. Пополнение системы осуществляется как автоматически (путем анализа научных публикаций), так и исследователями: депонирование данных секвенирования в открытых архивах (SRA, ENA и др.) требует обязательного предоставления метаинформации об экспериментах и биологических образцах. Ключевое преимущество Entrez – наличие программного интерфейса Entrez Programming Utilities (E-utilities) (Sayers, 2022), позволяющего осуществлять расширенный поиск и автоматизированное извлечение данных.

В задачи исследователя довольно часто входит поиск существующих наборов сырых данных, полученных в экспериментах по определенной тематике. Комбинация программного доступа к ресурсам Entrez с современными NLP-алгоритмами для анализа биомедицинских текстов позволяет создать унифицированную систему эффективного отбора интересующих данных (включая исходные данные секвенирования и сопроводительную метаинформацию). Существующие решения, например iSeq (Chao et al., 2024) и SampleExplorer (Chin, Lassmann, 2024), не обеспечивают универсального подхода и имеют ограниченную функциональность. В настоящей работе представлена система Alembic, предназначенная для анализа данных в открытых репозиториях NCBI и структурированного извлечения нужной информации. Название системы отражает сущность процесса обработки данных (alembic, аламбик – тип перегонного куба, предназначенного для извлечения полезных компонентов из больших объемов сырья).

Материалы и методы

В основе Alembic лежит клиент-серверная архитектура. Клиентский модуль управляет формированием запросов и отвечает за визуализацию результатов, тогда как серверная часть обрабатывает запросы и выполняет извлечение

именованных сущностей. Общая схема модулей системы и процессов обработки приведена на рис. 1.

Система Alembic использует данные из архива коротких прочтений NCBI (SRA), доступные посредством открытого API Entrez. Поиск поддерживает практически любые термины: названия генов, организмов, биологических процессов, метаболитов и других сущностей. Результаты запроса обрабатываются модулем AlembicDump следующим образом: извлекаются универсальные идентификаторы NCBI (UID), на их основе загружаются метаданные в XML-формате, которые затем стандартизируются и преобразуются в таблицу с заданной структурой столбцов. Метаданные представлены в стандарте Entrez SRA, включающем 21 поле, которые могут содержать как свободный неформатированный текст (например, Study Abstract), так и набор идентификаторов заданного формата (Sample Accession и др.).

На следующем этапе полученные структурированные данные обрабатываются алгоритмом AlembicNLP для извлечения именованных сущностей с использованием предобученных биомедицинских языковых моделей. Сначала входной текст последовательно проходит предобработку с помощью sciSpaCy (версия 2.0.18): преобразование в нижний регистр, очистку от специальных символов, пунктуации и изолированных цифр, фильтрацию стоп-слов (по словарю NLTK, www.ntlk.org, версия 3.8.1), лемматизацию и токенизацию. После чего трансформированный текст анализируется биомедицинской моделью AIONER (Luo et al., 2023) на основе архитектуры Bioformer, которая идентифицирует фрагменты текста, соответствующие биомедицинским сущностям. Исходный код AIONER, находящийся в открытом доступе, встроен в Alembic и не является зависимым элементом, требующим дополнительной установки.

В Alembic используются модели AIONER bioformer-cased-v1.0 и BiomedNLP-PubMedBERT-base-uncased-abstract, предобученные на данных PubMed и имеющиеся в открытом доступе (https://huggingface.co/lingbionlp/AIONER-0415/tree/main). Эти модели обучались на информации, имеющейся в описании статей (abstracts), депонированных в каталоге PubMed (Luo et al., 2023). После распознавания полученные элементы автоматически классифицируются по установленным категориям – гены (Gene), болезни (Disease), виды (Species), клеточные линии (Cell Line), генетические варианты (Variant) и химические соединения (Chemical). Финальный результат работы алгоритма AlembicNLP представляет собой аннотированный список сущностей с указанием их типа и позиционных координат в исходном тексте.

Клиентская часть приложения реализована на базе фреймворка Vite JS (версия 7.0), обеспечивающего быструю сборку и высокую производительность разработки. Приложение построено с помощью библиотеки React версии 19. Для создания интерфейса применяются компоненты из библиотеки Material UI (MUI) версии 7, а также специализированные компоненты для работы с таблицами данных (MUI X DataGrid). Взаимодействие с

Рис. 1. Схема пакета программ Alembic, включающего модули AlembicDump (предобработка результатов поиска) и AlembicNLP (извлечение именованных сущностей).

сервером осуществляется через асинхронные запросы с использованием библиотеки Axios (версия 1.6.7).

Функционал приложения реализован в виде двух взаимосвязанных модулей. Первый модуль предоставляет интерфейс поиска биомедицинскх терминов с последующим отображением результатов в виде агрегированной таблицы. Табличное представление результатов поиска по ключевым словам оптимизировано для работы с большими объемами данных.

Второй модуль отвечает за подготовку данных для ИИ-обработки. После выбора пользователем нужных ему полей система автоматически извлекает ключевые термины, классифицируя их по заданным категориям. Обработка выполняется серверной частью, тогда как клиентский интерфейс обеспечивает структурированное представление результатов в удобном формате. Такая архитектура минимизирует когнитивную нагрузку на пользователя и существенно упрощает анализ сложных текстовых описаний условий экспериментов.

Визуализация результатов группировки оснащена интерактивными инструментами: фильтрами терминов, быстрым поиском и динамической подсветкой релевантных элементов, что позволяет эффективно анализировать большие объемы данных, избавляя пользователя от необходимости изучения полных описаний экспериментов. Интерфейс адаптивен и оптимизирован для всех типов устройств – от десктопов до мобильных платформ.

Типичный сценарий работы с системой Alembic включает три стадии. В качестве примера рассмотрим поиск транскриптомных данных, полученных в ходе экспериментов по изучению болезни Альцгеймера. На первом этапе пользователь проводит поиск экспериментов в базе данных NCBI SRA с помощью набора ключевых слов, например названия генов, заболеваний, протоколов секвенирования и т. д., используя окно поиска (рис. 2). В данном случае это “Alzheimer”, “disease” и “RNASeq”. Система начинает поиск и показывает число найденных экспериментов (130 в приведенном примере), имеющихся в базе SRA. В случае когда число найденных экспериментов превы-

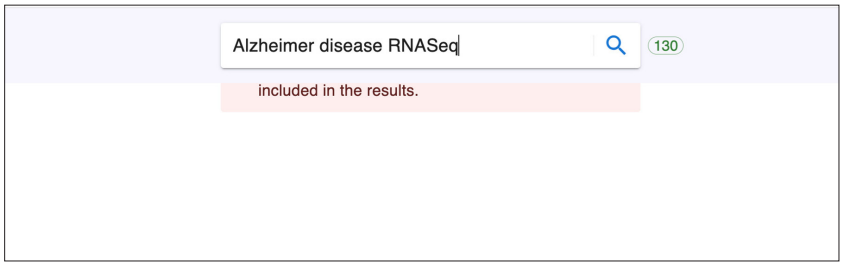

Рис. 2. Окно поиска системы Alembic.

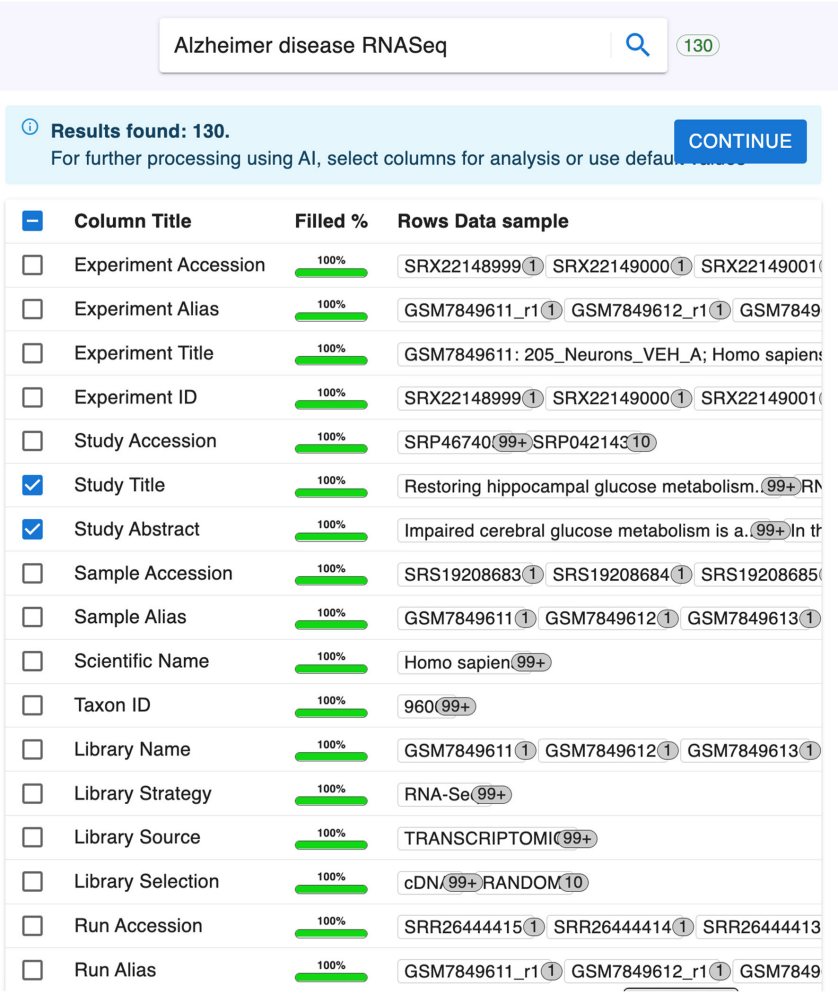

Рис. 3. Табличное представление результатов поиска по ключевым словам.

шает 1000 шт., выводится предупреждение об этом и исследователь может либо изменить термины, по которым ведется поиск, либо использовать первые 1000 из найденных экспериментов (сценарий по умолчанию).

На втором этапе Alembic структурирует полученные метаданные и представляет их в табличном виде (рис. 3). Формат включает в себя названия стандартных полей метаданных EntRez (Column Title), долю в процентах найденных экспериментов, имеющих в своих метаданных такое поле (Filled %) и примеры метаданных (Rows Data sample). Используя флажки в крайней левой колонке, пользователь может выбирать несколько типов метаданных для дальнейшего анализа, включая обработку полей данных средствами искусственного интеллекта.

На третьем этапе, нажав кнопку CONTINUE, пользователь запускает NLP-обработку и получает окончательный результат (рис. 4). В данном примере приведена визуализация результатов поиска сущностей в формате «облако слов», wordCloud. В правом верхнем углу окна приведены ссылки для загрузки файлов с результатами в текстовом формате, с полями, разделенными табуляцией (tab-separated text, tsv), как до обработки ИИ – RESULTS.TSV,

Рис. 4. Визуализация результатов обработки метаданных PubMedBERT.
Найденные именованные сущности представлены в виде облака слов (wordcloud).

так и после обработки – AI_PROCESSED.TSV. Полученные форматированные файлы могут быть напрямую импортированы в реляционные базы данных на основе sqLite, MySql или Postgress SQL. Наконец, файл DOWNLOADS.TXT содержит прямые ссылки для скачивания сырых данных секвенирования из репозитория SRA. В нижней части страницы приводятся список экспериментов и для каждого из них – список образцов, для которых получены данные секвенирования, доступные для скачивания.

Внешний интерфейс проекта реализован с использованием модулей JavaScript. Архитектура спроектирована с учетом масштабируемости; добавление новых функциональных модулей не требует структурных изменений. Исходный код проекта и полная документация к нему находятся в репозитории открытого доступа GitHub (лицензия MIT): https://github.com/shitohana/Alembic.

Для установки Alembic на компьютере пользователя необходимо наличие в системе языка python версии не ниже 3.12 и платформы docker версии не ниже 24.0.2. Остальные пакеты, необходимые для функционирования, в том числе библиотеки python TensorFlow (версия 2.3.0), Transformer (версия 4.18.0) и stanza (версия 1.4.0), необходимые для функционирования AIONER, устанавливаются автоматически. После установки пакета взаимодействие с системой осуществляется через локальный веб-интерфейс.

Результаты и обсуждение

Обработка больших объемов биомедицинских текстов потребовала создания многочисленных систем для распознавания именованных сущностей. Наиболее развитые решения в этой области – scispaCy и AIONER. scispaCy предлагает модели, предобученные на биомедицинских корпусах (GENIA, MedMentions), которые позволяют эффективно решать ключевые задачи NLP: токенизацию, POS-разметку (part-of-speech – разметка по частям речи) – технику в обработке естественного языка (NLP), присваивающую грамматическую категорию (например, существительное, глагол или прилагательное) каждому слову в тексте, синтаксический анализ и NER, учитывая при этом биомедицинскую специфику текста.

В отличие от этого, AIONER специализируется исключительно на BioNER, реализуя инновационный подход All-in-One (AIO), сущность которого заключается в одновременном распознавании нескольких терминов из различных наборов данных, применяемых для обучения. По сравнению с другими алгоритмами, где разметка текста по токенам происходит отдельно для каждой категории (например, категорий Gene, Disease), AIO обрабатывает все категории одновременно, используя специальные метки (tags), для того чтобы избежать ошибок при атрибутировании отдельных слов или их комбинаций. Данная архитектура повышает точность токенизации и, таким образом, решает проблему переобучения и слабой обобщающей способности моделей, вызванную нехваткой размеченных биомедицинских данных. В основе AIONER лежат современные предобученные языковые модели (PubMedBERT и аналоги), дополненные слоем условных случайных полей (conditional random fields, CRF) для точного определения границ сущностей. Благодаря этим преимуществам AIONER выбран базовым алгоритмом для пакета Alembic.

Растущий объем биомедицинских текстов стимулировал разработку многочисленных систем извлечения именованных сущностей (named entity recognition, NER), ведущими примерами которых являются scispaCy и AIONER. scispaCy предлагает модели, предварительно обученные на биомедицинских корпусах, таких как GENIA и MedMentions. Эти модели эффективно выполняют основные задачи обработки естественного языка, а именно

токенизацию, разметку частей речи (part-of-speech, POS), синтаксический анализ и NER, одновременно учитывая доменно-специфические лингвистические особенности. Под разметкой частей речи здесь понимается фундаментальный метод NLP, заключающийся в присвоении каждому слову в тексте грамматической категории (например, существительное, глагол, прилагательное).

Среди существующих решений для извлечения структурированных данных из открытых репозиториев выделяются iSeq и SampleExplorer, функциональность которых принципиально отличается от NLP-подхода Alembic. iSeq функционирует как инструмент командной строки для пакетной загрузки NGS-данных по известным идентификаторам из репозиториев GSA/SRA/ENA/DDBJ. Он автоматизирует скачивание с проверкой целостности, но полностью исключает семантический анализ метаданных, работая только с предопределенными списками идентификаторов. SampleExplorer, напротив, применяет языковые модели для поиска транскриптомных данных в архивах типа ARCHS4. Его алгоритм сочетает векторное представление текстовых метаданных с транскриптомным сходством, выявляя семантически близкие эксперименты по генам или описаниям. Однако инструмент остается ориентированным на поиск аналогичных образцов, а не на детальную аннотацию сущностей внутри описаний конкретных экспериментов.

В отличие от iSeq и SampleExplorer, Alembic обеспечивает принципиально иной уровень обработки данных: пакет выполняет глубинный NLP-анализ биомедицинских текстов, включая метаданные экспериментов, с применением передовых моделей – BERT, BioBERT, scispaCy и AIONER. Ключевое преимущество Alembic заключается в использовании PubMedBERT, дообученного на обширных биомедицинских корпусах (PubMed, PMC). Эта модель демонстрирует превосходство как над базовым BERT, так и над классическими подходами, достигая исключительной точности в распознавании сложной биомедицинской терминологии и контекстуальных зависимостей за счет платформы AIONER, составляющей ядро Alembic (Luo et al., 2023).

Заключение

Система Alembic реализует принципиально новый подход к анализу биомедицинских данных открытых репозиториев через глубинное NLP-структурирование информации. Alembic генерирует готовые к импорту выгрузки метаданных (совместимые с SQLite/PostgreSQL) и автоматизирует получение соответствующих сырых данных секвенирования. Главное преимущество решения заключается в создании специализированных, проектно-ориентированных, локальных баз данных. Используемая в Alembic платформа AIONER превосходит аналогичные инструменты благодаря комплексной обработке текстов на основе современных языковых моделей, обеспечивающей точную категоризацию биомедицинских сущностей. Интуитивный интерфейс Alembic с интерактивной визуализацией результатов является принципиальным упрощением по сравнению с инструментами поиска, использующими командную строку, облегчая рутинный поиск данных.

Таким образом, пакет Alembic существенно облегчает операции по извлечению информации из репозиториев открытого доступа, преобразуя неструктурированные метаданные в формализованное машинно-читаемое знание.

Список литературы / References

Aronson A.R., Lang F.M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229-236. doi 10.1136/jamia.2009.002733

Chao H., Li Z., Chen D., Chen M. iSeq: an integrated tool to fetch public sequencing data. *Bioinformatics*. 2024;40(11):btae641. doi 10.1093/bioinformatics/btae641

Chin W.L., Lassmann T. SampleExplorer: using language models to discover relevant transcriptome data. *Bioinformatics*. 2024;41(1):btae759. doi 10.1093/bioinformatics/btae759

Devlin J., Chang M.W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv.* 2019. doi 10.48550/arXiv.1810.04805

Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234-1240. doi 10.1093/bioinformatics/btz682

Luo L., Wei C.-H., Lai P.-T., Leaman R., Chen Q., Lu Z. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics.* 2023;39(5):btad310. doi 10.1093/bioinformatics/btad310

Neumann M., King D., Beltagy I., Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, 2019;319-327. doi 10.18653/v1/W19-5034

Sayers E. The E-utilities in-depth: parameters, syntax and more. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US), 2022. Available at: https://www.ncbi.nlm.nih.gov/books/NBK25499/. Accessed: Jul. 30, 2025

Wang X., Zhang Y., Ren X., Zhang Y., Zitnik M., Shang J., Langlotz C., Han J. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*. 2019;35(10):1745-1752. doi 10.1093/bioinformatics/bty869

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 06.08.2025. После доработки 21.10.2025. Принята к публикации 06.11.2025.