Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2024;28(8):940-949

doi 10.18699/vjgb-24-101

Онтологии в моделировании и анализе больших генетических данных

Н.Л. Подколодный [1, 2, 3, 4], О.А. Подколодная [1], В.А. Иванисенко [1, 4], М.А. Марченко [2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

pnl@bionet.nsc.ru

Аннотация. Для систематизации и эффективного использования огромного объема экспериментальных данных, накопленных в области биоинформатики и биомедицины, необходимы новые подходы, основанные на онтологиях, включая автоматизированные методы семантической интеграции гетерогенных экспериментальных данных, методы создания больших баз знаний и самоинтерпретируемые методы анализа больших разнородных данных на основе глубокого обучения. В статье кратко представлены особенности предметной области (биоинформатика, системная биология, биомедицина), формальные определения понятия онтологии и графов знаний, приведены примеры применения онтологий для семантической интеграции гетерогенных данных и создания больших баз знаний, а также интерпретации результатов глубокого обучения на больших данных. В качестве примера успешного проекта описана база знаний Gene Ontology, которая помимо терминологических знаний и аннотаций генов (GOA) включает модели причинных влияний (GO-CAM). Это делает ее полезной не только для геномной биологии, но и для системной биологии, а также для интерпретации крупномасштабных экспериментальных данных. Обсуждается подход к созданию больших онтологий с использованием шаблонов проектирования на примере онтологии биологических атрибутов (OBA). Здесь большая часть классификации автоматически вычисляется на основе ранее созданных эталонных онтологий с помощью автоматизированного логического вывода, за исключением небольшого числа высокоуровневых понятий. Одной из основных проблем глубокого обучения является отсутствие интерпретируемости, поскольку нейронные сети часто функционируют как «черные ящики», не способные объяснить свои решения. В нашей статье описаны подходы к созданию методов интерпретации моделей глубокого обучения и представлены два примера самообъясняемых моделей глубокого обучения на основе онтологий. Модель Deep GONet, которая интегрирует Gene Ontology в иерархическую архитектуру нейронной сети, где каждый нейрон представляет биологическую функцию. Эксперименты с наборами данных диагностики рака показывают, что Deep GONet легко интерпретируется и обладает высокой производительностью для различения раковых и нераковых образцов. Модель ONN4MST, использующая онтологии биома для отслеживания микробных источников образцов, ниши которых ранее были мало изучены или неизвестны, и обнаружения микробных загрязнителей. ONN4MST может отличать образцы от онтологически близких биомов и, таким образом, предлагает количественный способ охарактеризовать развитие микробного сообщества кишечника человека. Оба примера демонстрируют высокую производительность и интерпретируемость, что делает их ценными инструментами для анализа и интерпретации больших данных в биологии.

Ключевые слова: онтологии; биоинформатика; системная биология; анализ больших данных; глубокое обучение; интерпретируемость.

Для цитирования: Подколодный Н.Л., Подколодная О.А., Иванисенко В.А., Марченко М.А. Онтологии в моделировании и анализе больших генетических данных. *Вавиловский журнал генетики и селекции*. 2024;28(8):940-949. doi 10.18699/vjgb-24-101

Финансирование. Работа выполнена при поддержке бюджетных проектов FWNR-2022-0020 и FWNM-2022-0005.

Ontologies in modelling and analysing of big genetic data

N.L. Podkolodnyy [1, 2, 3, 4], O.A. Podkolodnaya [1], V.A. Ivanisenko [1, 4], M.A. Marchenko [2, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia

pnl@bionet.nsc.ru

Abstract. To systematize and effectively use the huge volume of experimental data accumulated in the field of bioinformatics and biomedicine, new approaches based on ontologies are needed, including automated methods for semantic integration of heterogeneous experimental data, methods for creating large knowledge bases and self-interpreting methods for analyzing large heterogeneous data based on deep learning. The article briefly presents the

features of the subject area (bioinformatics, systems biology, biomedicine), formal definitions of the concept of ontology and knowledge graphs, as well as examples of using ontologies for semantic integration of heterogeneous data and creating large knowledge bases, as well as interpreting the results of deep learning on big data. As an example of a successful project, the Gene Ontology knowledge base is described, which not only includes terminological knowledge and gene ontology annotations (GOA), but also causal influence models (GO-CAM). This makes it useful not only for genomic biology, but also for systems biology, as well as for interpreting large-scale experimental data. An approach to building large ontologies using design patterns is discussed, using the ontology of biological attributes (OBA) as an example. Here, most of the classification is automatically computed based on previously created reference ontologies using automated inference, except for a small number of high-level concepts. One of the main problems of deep learning is the lack of interpretability, since neural networks often function as "black boxes" unable to explain their decisions. This paper describes approaches to creating methods for interpreting deep learning models and presents two examples of self-explanatory ontology-based deep learning models: (1) Deep GONet, which integrates Gene Ontology into a hierarchical neural network architecture, where each neuron represents a biological function. Experiments on cancer diagnostic datasets show that Deep GONet is easily interpretable and has high performance in distinguishing cancerous and non-cancerous samples. (2) ONN4MST, which uses biome ontologies to trace microbial sources of samples whose niches were previously poorly studied or unknown, detecting microbial contaminants. ONN4MST can distinguish samples from ontologically similar biomes, thus offering a quantitative way to characterize the evolution of the human gut microbial community. Both examples demonstrate high performance and interpretability, making them valuable tools for analyzing and interpreting big data in biology.

Key words: ontologies; big data analysis; bioinformatics; systems biology; deep learning; interpretability.

For citation: Podkolodnyy N.L., Podkolodnaya O.A., Ivanisenko V.A., Marchenko M.A. Ontologies in modelling and analysing of big genetic data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8): 940-949. doi 10.18699/vjgb-24-101

Введение

Термин «большие данные» (Big Data) обозначает объемные наборы данных, которые отличаются значительными размерами, разнообразием и сложностью, что делает их трудными для обработки и анализа с помощью традиционных методов. Кроме того, такие данные часто имеют неполноту и неопределенность, что усложняет задачу контроля их качества и точности (Qaiser, Ghulam, 2023).

Появление качественно новых возможностей для проведения исследований, основанных на высокопроизводительных экспериментальных технологиях, таких как массовое параллельное секвенирование ДНК, многолокусное генотипирование, многопараметрическое профилирование экспрессии генов с использованием ДНК-чипов, технологии ChIP-on-chip, а также протеомные и метаболомные технологии, привело к накоплению беспрецедентно больших объемов экспериментальных данных и знаний (Stephens et al., 2015). Гетерогенность молекулярно-биологической информации и ее сложность затрудняют анализ, систематизацию и применение этих данных для решения конкретных задач в биоинформатике, биотехнологии, фармакологии и персонализированной медицине.

Для освоения, систематизации и эффективного использования огромного объема данных необходимы новые подходы к их обработке. В частности, это автоматизированные методы семантической интеграции гетерогенных данных, одним из ключевых этапов которой является согласование понятий предметной области, а также способов их описания и использования. Согласованное описание конкретной предметной области называется онтологией.

Онтологии позволяют представлять понятия в формате, пригодном для машинной обработки, и выступают в роли посредника между пользователем и информационной системой, а также между членами научного сообщества при обмене данными. Таким образом, онтологии становятся важным инструментом в биоинформатике и системной биологии, способствуя семантической интеграции экспериментальных данных и знаний с целью создания «единой картины мира». Кроме того, они помогают решать проблемы, возникающие при анализе больших данных, преодолевая разнородность и недостатки качества данных, а также улучшая интерпретацию результатов глубокого обучения. Онтологии повышают масштабируемость и эффективность обработки больших объемов информации, что делает их незаменимыми в современных научных исследованиях.

Ранее в обзоре (Podkolodnyy et al., 2016) были представлены примеры онтологий, описывающих биологические системы на различных уровнях организации живых систем. В настоящей статье будут рассмотрены примеры применения онтологий для интеграции гетерогенных данных и создания больших баз знаний, а также интерпретации результатов анализа данных.

Формальное представление онтологий

В информатике термин «онтология» обозначает концептуальную модель, которая описывает объекты, их свойства и взаимосвязи между ними (Chandrasekaran et al., 1999). Онтология включает в себя набор понятий (терминов) конкретной предметной области и их определения, а также всю информацию, связанную с этими понятиями, такую как свойства, отношения, ограничения, аксиомы и утверждения. Эта информация необходима для описания и решения задач в выбранной предметной области (Podkolodnyy et al., 2016).

Таким образом, формальная модель онтологии может быть представлена в виде упорядоченной тройки конечных множеств O = <T, R, F>, где T – это конечное и непустое множество классов и концептов (понятий, терминов) предметной области, которая рассматривается в определенном контексте (в нашем случае: биоинформатика, системная биология, биотехнология и биомедицина);

R – конечное множество отношений между концептами данной предметной области; F – конечное множество функций интерпретации, заданных на понятиях и/или отношениях онтологии O, а также аксиом, используемых для моделирования утверждений, которые всегда являются истинными. Это ограничивает интерпретацию и обеспечивает корректное использование понятий.

Одним из наиболее эффективных подходов к описанию и использованию знаний о предметной области являются дескриптивные логики, которые определяют формальный язык для описания понятий (концептов, классов, категорий или сущностей) и отношений между ними (называемых ролями), а также для формулирования утверждений о фактах и запросов к ним, включая проверку выполнимости и включения. Кроме того, дескриптивная логика включает конструкторы (операции) для создания понятийных выражений, такие как конъюнкция, дизъюнкция и определение отношений.

В базе знаний предметной области с точки зрения дескриптивной логики можно выделить две основные категории знаний. Первая категория включает общие знания о множестве классов понятий, их свойствах и отношениях между ними, что обозначается как терминологические знания (terminological knowledge), или T-Box. Вторая категория охватывает знания об индивидуальных объектах (экземплярах классов), их свойствах и связях с другими объектами, известные как утверждающие знания (assertional knowledge), или A-Box. Таким образом, T-Box описывает предметную область на уровне абстрактных понятий, в то время как A-Box является базой данных. Важно отметить, что обе компоненты базы знаний взаимосвязаны и дополняют друг друга.

Для систематического моделирования сложных систем, организмов и заболеваний, а также для представления знаний в области биоинформатики и системной биологии часто применяются графы знаний (ГЗ). В соответствии с определением, приведенным в работе (Callahan et al., 2024), граф знаний является структурой данных, которая отображает множество разнородных сущностей и различные типы отношений между ними. Эта структура служит абстрактной основой, способной генерировать новые знания, а также выявлять и разрешать расхождения или противоречия, что делает ее полезной для решения различных задач и применения в различных сценариях.

Можно выделить три типа графов знаний в зависимости от сложности представления и функциональности использования.

Простые графы – это наиболее распространенный и базовый тип графов. В таких графах сущности представлены в виде узлов, а ребра описывают отношения между ними. Обычно в простых графах отсутствует формальная семантика для ребер и узлов, что делает их простыми в применении, но ограничивает возможности для более глубокого анализа и интерпретации данных.

Гибридный граф, или граф свойств. Гибридные графы предназначены для моделирования сущностей и их отношений с использованием сочетания стандартных представлений сети и формальной семантики, таких как Resource Description Framework (RDF: https://www.w3.org/RDF) и RDF Schema (RDFS: https://www.w3.org/TR/rdf11-mt). В отличие от простых графов, гибридные графы, основанные на этих стандартах, облегчают интеграцию с другими ресурсами и обеспечивают большую возможность для автоматизированного вывода знаний. Это делает их более мощным инструментом для представления и обработки сложной информации.

Сложный граф, такой как в системе KaBOB (Livingston et al., 2015; Podkolodnyy et al., 2016), часто строится на основе Web Ontology Language (OWL). Сложные графы обладают высокой выразительностью, что позволяет эффективно генерировать новые знания с помощью дедуктивного вывода (Подколодный и др., 2012). Благодаря явной семантике, OWL предоставляет значительные преимущества в интеграции больших объемов биомедицинских данных по сравнению с RDF/RDFS, что делает его особенно полезным для сложных задач в области биоинформатики и системной биологии.

В качестве примера рассмотрим сеть основных взаимосвязанных биомедицинских концепций, необходимых для моделирования знаний о путях, генетических вариантах, заболеваниях и фармацевтических методах лечения (рис. 1). На верхнем уровне представлены анатомические сущности, такие как ткани, клетки и биологические жидкости (компартменты), содержащие геномные сущности (ДНК, РНК, мРНК и белки). ДНК кодирует гены, транскрибирующиеся в мРНК и транслирующиеся в белки, которые имеют молекулярные функции, могут взаимодействовать друг с другом и участвовать в путях и биологических процессах.

В последнее время было разработано несколько программных систем, таких как KG-HUB (Caufield et al., 2023), Clinical KG (CKG) (Santos et al., 2022), RTX-KG2 (Wood et al., 2022), BioCypher (Lobentanzer et al., 2023) и Knowledge Base Of Biomedicine (KaBOB) (Livingston et al., 2015; Podkolodnyy et al., 2016), которые обеспечивают широкие функциональные возможности создания и использования графов знаний в биоинформатике и биомедицине, включая интеграцию больших гетерогенных данных.

В работе (Callahan et al., 2024) описана семантическая экосистема PheKnowLator (Phenotype Knowledge Translator) для автоматизации построения онтологических ГЗ с полностью настраиваемым представлением знаний. Экосистема включает в себя различные компоненты для создания и использования ГЗ для решения различных прикладных задач, а также предварительно построенные графы знаний.

Интеграция больших данных и создание баз знаний на основе онтологий

В настоящее время в области биоинформатики, системной биологии, агробиологии, биомедицины разработано более тысячи онтологий, которые можно использовать для описания и интеграции знаний, анализа данных, а также вывода новых знаний (https://bioportal.bioontology.org/ontologies).

В качестве примера одного из самых успешных проектов создания онтологий можно привести проект Gene Ontology (GO) (http://www.geneontology.org/). В GO описаны

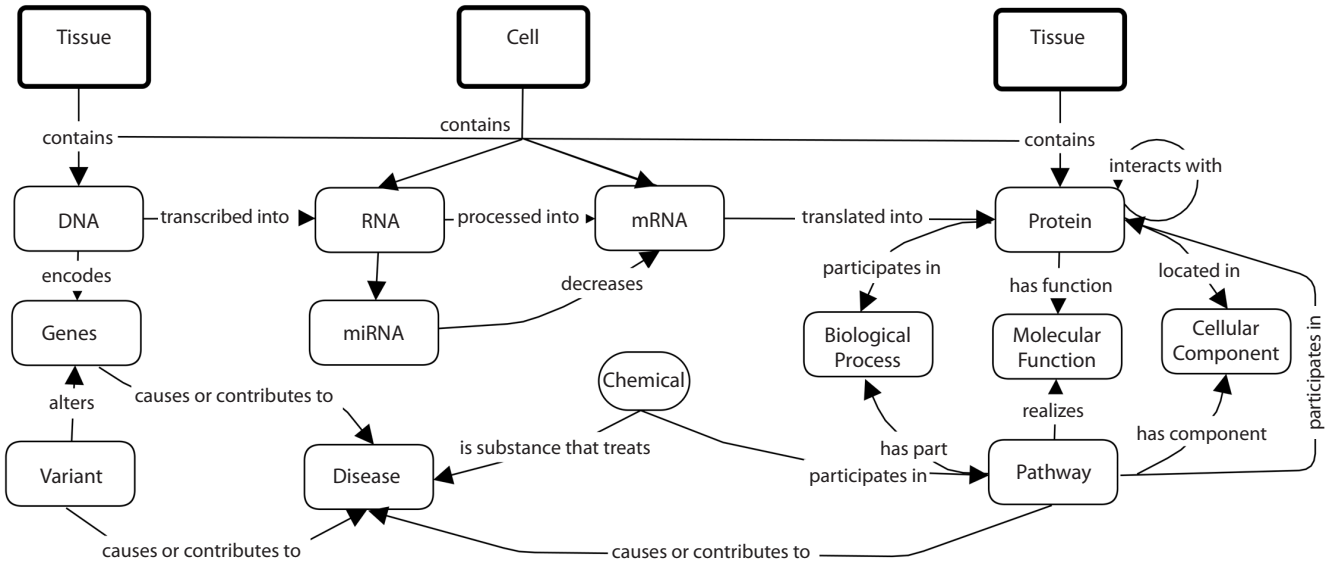

Рис. 1. Представление знаний об уровнях биологической организации, лежащих в основе описания заболеваний человека (Callahan et al., 2024).

текущие знания о типах функциональных характеристик (всего более 40 тыс. понятий), которыми может обладать генный продукт. В состав GO входят три раздела:

- Молекулярная функция (Molecular function) – элементарная молекулярная активность или роль, которую может выполнять ген, продукт гена в каких-либо биологических процессах. Всего описано 10365 терминов (https://geneontology.org/stats.html, дата обращения 08.09.2024).
- Биологические процессы (Biological process) – «биологическая программа», включающая набор молекулярных событий или активностей, действующих согласованно для достижения определенного результата и относящихся к функционированию интегрированных живых единиц: клеток, ткани, органов и организмов. В отличие от функции, процесс должен иметь несколько различающихся этапов с определенным началом и концом. Всего описано 26552 термина (https://geneontology.org/stats.html, дата обращения 08.09.2024).
- Клеточные компоненты (Cellular component) – часть анатомической структуры, в которой описывается локализация гена или его продукта в организме на уровнях клеточных структур и макромолекулярных комплексов или групп продуктов генов. Всего описано 4022 термина (https://geneontology.org/stats.html, дата обращения 08.09.2024).

Основные отношения между понятиями, которые используются в GO, включают простое отношение класс-подкласс (is_a), отношение часть-целое (part_of), отношения regulates, positively_regulates и negatively_regulates, которые описывают отношения между биологическими процессами, молекулярными функциями или биологическими свойствами. Свойство транзитивности отношений, используемых в GO, позволяет строить решетку отношений между понятиями и выполнять логический вывод о свойствах понятий и их отношений (Podkolodnyy et al., 2016).

На основе GO создана база знаний, которая кроме терминологических знаний (онтологии генов GO) включает результаты аннотации генов GOA (Gene Ontology Annotation, http://www.ebi.ac.uk/GOA), т. е. знания об индивидуальных объектах – генах и их продуктах (Huntley et al., 2015). В настоящее время GOA включает более 7.6 млн GO аннотаций для почти 1.54 млн белков и более 4.4 тыс. видов организмов.

Исходно на ранней стадии развития GO аннотация гена или его продукта (белка или РНК) осуществлялась независимо по молекулярным функциям, биологическим процессам или клеточным компонентам. Для того чтобы получить информацию о том, какую функцию выполняет ген или его продукт (РНК, белок) в том или ином биологическом процессе и той или иной клеточной структуре, понадобилась разработка еще одной компоненты базы знаний GO – модели причинных влияний между генными продуктами GO-CAM (Thomas et al., 2019).

GO-CAM связывает несколько аннотаций GO вместе, чтобы создать модели биологических процессов, соединяющие активности более чем одного генного продукта вместе в причинно-следственные сети и позволяющие специфицировать биологический контекст (например, тип клетки/ткани), в котором происходят активности. В качестве примера можно привести представление одной и той же биологической модели, описывающей, как убиквитин-протеинлигаза E3 NEDD4 подавляет транскрипцию РНК в ответ на повреждение ДНК, вызванное УФ-излучением, двумя способами: как набор разрозненных аннотаций GO, каждая из которых охватывает частичное описание общей функции (рис. 2, *а*), и как схема GO-CAM, связывающая аннотации GO в структурированную модель функций NEDD4, включая влияние активности NEDD4 на активность макромолекулярного комплекса РНК-полимеразы II (см. рис. 2, *б*).

Базовой единицей GO-CAM является единица активности продукта гена, которая объединяет аннотацию GO MF

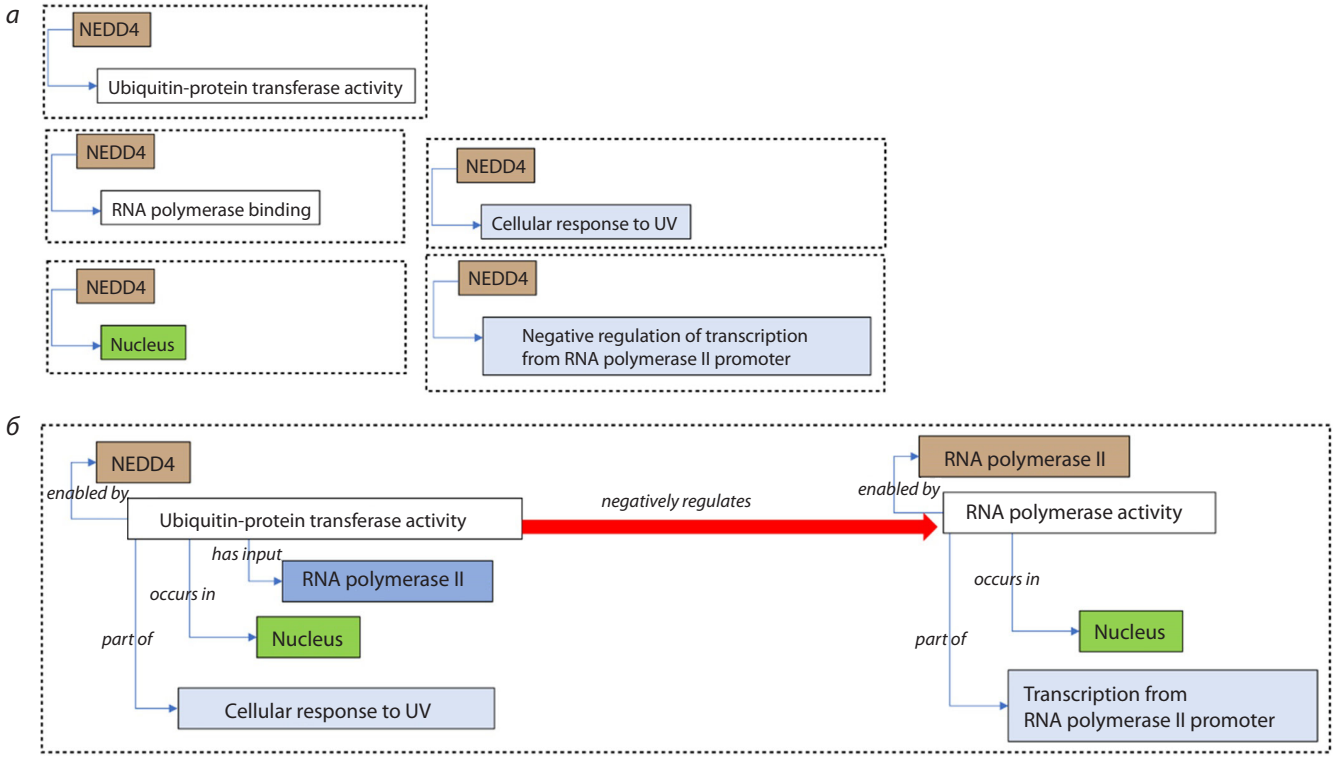

Рис. 2. Одна и та же биологическая модель, описывающая двумя способами, как NEDD4 подавляет транскрипцию РНК в ответ на повреждение ДНК, вызванное УФ-излучением: *а* – как набор разрозненных аннотаций GO, каждая из которых охватывает частичное описание общей функции; *б* – как схема GO-CAM, связывающая аннотации GO в структурированную модель функций NEDD4, включая влияние активности NEDD4 на активность макромолекулярного комплекса РНК-полимеразы II (Thomas et al., 2019).

(молекулярная активность) вместе с аннотациями GO CC (клеточные компоненты) и GO BP (биологический процесс), обеспечивающими биологический контекст активности. Контекст может быть дополнительно задан другими онтологиями, включая онтологии типа клетки (Cell Type Ontology) (Diehl et al., 2016), тканевого/анатомического расположения (с использованием нескольких различных онтологий в зависимости от вида организма, например, интегрированная онтология межвидовой анатомии, охватывающая животных и объединяющая несколько видоспецифических онтологий – Uberon (https://obophenotype.github.io/uberon/) (Mungall et al., 2012)), или неживотные онтологии, такие как Plant ontology (https://planteome.org/) (Cooper, Jaiswal, 2016), или описание временного периода (например, биологическая фаза GO). Единицы активности связаны между собой причинно-следственными связями из Онтологии Отношений (Smith et al., 2005).

Причинно-следственные сети в моделях GO-CAM позволяют использовать совершенно новые приложения, такие как сетевой анализ геномных данных и логическое моделирование биологических систем. Кроме того, модели могут оказаться полезными для визуализации пути. Например, основанное на активности представление GO-CAM совместимо с «диаграммами потока активности» стандарта Systems Biology Graphical Notation (SBGN) (Bergmann et al., 2020).

Таким образом, GO-CAM дает возможность использовать массивную базу знаний GO и GOA, собранную за последние двадцать лет, в качестве основы не только для геномной биологии представления функции гена, но и для более обширного представления системной биологии и его новых приложений к интерпретации крупномасштабных экспериментальных данных.

Пример GO анализа генов ассоциативной генной сети ревматоидного артрита

Ранее в Институте цитологии и генетики СО РАН была разработана программно-информационная система ANDSystem для автоматизированного извлечения медико-биологических знаний из научных публикаций и большого числа биологических и биомедицинских фактографических баз данных (Ivanisenko et al., 2015, 2019). База знаний ANDSystem представляет собой уникальный ресурс, содержащий формализованную информацию в виде ассоциативных генных сетей (графов знаний) о почти 44 млн взаимодействий различных типов между молекулярно-генетическими объектами.

Оригинальная онтология, лежащая в основе ANDSystem, дает очень подробное описание предметной области. В базе знаний ANDSystem описаны молекулярно-генетические объекты (белки, гены, метаболиты, микроРНК), биологические процессы, фенотипические признаки, лекарственные средства и их побочные эффекты, заболевания и т. д., а также более 25 типов взаимодействий между этими объектами, включая: физические взаимодействия с образованием молекулярных комплексов (белок/белок, белок/ДНК, метаболит/белок); каталитические реакции и протеолитические события с участием субстрата/фер-

мента/продукта; регуляторные взаимодействия; функции/активности, транспорта и стабильности белков; регуляцию трансляции белка с участием миРНК; регуляцию биологических процессов и фенотипических признаков с участием белков, метаболитов и лекарств; ассоциативные взаимодействия генов, белков, метаболитов, биологических процессов, фенотипических признаков с заболеваниями и т. д.

Примером типичной задачи использования ANDsystem является реконструкция ассоциативной генной сети (графа знаний) ревматоидного артрита (РА), содержащей 1025 генов/белков и более 20 тыс. взаимодействий между ними. Анализ перепредставленности терминов биологических процессов в Gene Ontology для множества генов ревматоидного артрита, выполненный с помощью системы DAVID (https://david.ncifcrf.gov/tools.jsp), выявил 376 биологических процессов, статистически значимо связанных с ревматоидным артритом (см. таблицу). Значения *p*-value вычислялись на основе гипергеометрического распределения. Для учета множественного тестирования использовалась поправка Бонферрони.

Рассмотрим более детально процесс иммунного ответа (GO:0006955~immune response). В Gene Ontology описано 420 генов, связанных с термином "GO:0006955~immune response". Из них 158 присутствуют в ассоциативной сети ревматоидного артрита (рис. 3). Вероятность выявления столь большого числа генов по случайным причинам очень низка (*p*-value с поправкой Бонферрони $<4.69 \cdot 10^{-79}$),

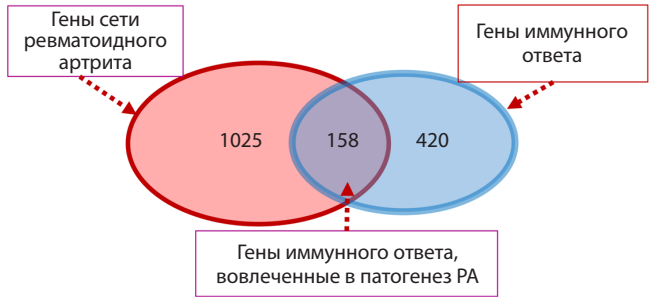

Рис. 3. Диаграмма Венна, описывающая пересечение генов сети ревматоидного артрита и генов иммунного ответа, связанных с термином "GO:0006955~immune response".

что указывает на высокую значимость взаимосвязи ревматоидного артрита с процессом иммунного ответа и на важнейшую роль иммунной системы в патогенезе этого заболевания.

Список первых из упорядоченных по статистической значимости (*p*-value с поправкой Бонферрони) биологических процессов, взаимосвязанных с ревматоидным артритом, представлен в таблице. Большая часть этих терминов так или иначе связана с процессами иммунного ответа и воспаления, играющими важную роль в патогенезе ревматоидного артрита. Эти процессы не являются независимыми.

Список первого 21 биологического процесса, статистически наиболее значимо связанного с ревматоидным артритом

Биологический процесс (Gene Ontology)	*p*-value с поправкой Бонферрони
GO:0006955~immune response	$4.69 \cdot 10^{-79}$
GO:0006954~ inflammatory response	$2.13 \cdot 10^{-70}$
GO:0060326~chemotaxis	$2.49 \cdot 10^{-30}$
GO:0007267~cell-cell signaling	$8.59 \cdot 10^{-28}$
GO:0032496~response to lipopolysaccharide	$7.41 \cdot 10^{-27}$
GO:0070098~chemokine-mediated signaling pathway	$3.91 \cdot 10^{-25}$
GO:1990256~signal transduction	$2.91 \cdot 10^{-24}$
GO:0071222~cellular response to lipopolysaccharide	$5.45 \cdot 10^{-24}$
GO:0050729~positive regulation of inflammatory response	$6.31 \cdot 10^{-24}$
GO:2001023~regulation of response to drug	$1.70 \cdot 10^{-23}$
GO:0070374~positive regulation of ERK1 and ERK2 cascade	$8.26 \cdot 10^{-23}$
GO:0001666~response to hypoxia	$9.11 \cdot 10^{-23}$
GO:0071864~positive regulation of cell proliferation	$2.52 \cdot 10^{-22}$
GO:0042102~positive regulation of T cell proliferation	$6.90 \cdot 10^{-22}$
GO:0045087~innate immune response	$2.09 \cdot 10^{-18}$
GO:0032729~positive regulation of interferon-gamma production	$2.38 \cdot 10^{-18}$
GO:0045766~positive regulation of angiogenesis	$2.90 \cdot 10^{-18}$
GO:0043066~negative regulation of apoptotic process	$5.72 \cdot 10^{-18}$
GO:0050731~positive regulation of peptidyl-tyrosine phosphorylation	$8.13 \cdot 10^{-18}$
GO:0007166~cell surface receptor signaling pathway	$8.40 \cdot 10^{-18}$
GO:0007568~aging	$1.28 \cdot 10^{-17}$

В частности, процесс иммунного ответа (термин “GO:0006955~immune response”) является обобщением таких терминов (см. таблицу), как “GO:0045087~innate immune response”, “GO:0032729~positive regulation of interferon-gamma production”, “GO:0060326~chemotaxis”, “GO:0042102~positive regulation of T cell proliferation”, “GO:1990256~signal transduction” и др.

Процесс воспалительного ответа (GO:0006954~inflammatory response) включает более частные процессы, описываемые терминами “GO:0032496~response to lipopolysaccharide”, “GO:0050729~positive regulation of inflammatory response”, “GO:1990256~signal transduction”, “GO:0001666~response to hypoxia”, “GO:0045766~positive regulation of angiogenesis”. И даже термин “GO:0007568~aging” связан с термином “GO:0006954~inflammatory response”, так как одним из механизмов старения являются хронические неинфекционные воспаления.

Приведенные на примере ревматоидного артрита результаты свидетельствуют, что подход к поиску генов, ассоциированных с конкретным заболеванием, с помощью ANDsystem и дальнейший GO анализ этой группы генов позволяет выявлять ключевые биологические процессы, участвующие в патогенезе данного заболевания.

Использование шаблонов проектирования онтологий для интеграции онтологий фенотипов и биологических атрибутов

Онтологии, имеющие логически богатую аксиоматизацию, обеспечивают такие мощные возможности, как автоматизированные рассуждения, классификацию и логические запросы. Однако ручное создание такого рода онтологий крайне затратно и требует от аннотаторов не только быть специалистами в предметной области, но и обладать знаниями в области логического моделирования (Slater et al., 2020).

Популярным подходом к решению данной проблемы является применение шаблонов проектирования и систем шаблонов для логических аксиом (Osumi-Sutherland et al., 2017). Это позволяет отделить курирование ссылочных терминов и их логических определений от их точной аксиоматической картины. Центральная идея состоит в том, чтобы задействовать небольшое количество шаблонов аксиом, реализующих шаблоны проектирования, которые могут быть созданы и поддерживаться экспертами по логике, и чтобы кураторы контента сосредоточились на выборе соответствующих терминов из различных ссылочных онтологий (например, терминов из онтологии Uberon для определения анатомических атрибутов).

Онтология биологических атрибутов (OBA) – это стандартизированная структура для наблюдаемых атрибутов, которые являются характеристиками организмов или частей организмов (Stefancsik et al., 2023). В отличие от большинства фенотипических онтологий, в OBA логические аксиомы определяют общие атрибуты без ссылки на какие-либо конкретные фенотипические изменения или состояния.

При создании OBA использовался шаблон проектирования типа Сущность-Качество (EQ), в котором фенотипическое качество (Q), такое как “height”, “mass” или “amount” из онтологии фенотипа и признаков (PATO) (Gkoutos et al., 2018), объединяется с сущностью (E), такой как анатомическая или химическая сущность, чтобы сформировать концепцию «биологического атрибута», называемого “trait”. Например, понятие “blood glucose amount” (OBA:VT0000188) включает класс “amount” (PATO:000070), который определяет характеристику глюкозы – “glucose” (CHEBI:17234) в крови – “blood” (UBERON:0000178).

В настоящее время в OBA применяется десять шаблонов признаков из Dead Simple Ontology Design Patterns (DOS-DP) (Osumi-Sutherland et al., 2017). Они были выбраны, потому что охватывают большинство анатомических, химических и клеточных атрибутов, являющихся центральными для интеграции данных геномики.

Богатая логическая аксиоматизация, основанная на шаблонах проектирования, необходима для обеспечения совместимости с существующими онтологиями фенотипов и другими типами данных, такими как анатомические, химические и биологические данные о метаболических путях и генных сетях.

Большинство атрибутов в OBA логически можно определить с помощью OWL, используя термины из соответствующих ссылочных онтологий, например Uberon (Mungall et al., 2012) или Chebi (Hastings et al., 2016). За исключением незначительного числа высокоуровневых понятий, большая часть классификации в OBA автоматически вычисляется на основе классификаций различных эталонных онтологий, с помощью автоматизированного логического вывода. Можно выделить два преимущества этого подхода: во-первых, не нужно вручную классифицировать какие-либо понятия, что значительно снижает стоимость кураторства классификации при одновременном повышении ее полноты. Во-вторых, многочисленные ссылки на ссылочные онтологии могут использоваться для самых разных целей, включая запрос (например, выбор всех данных, где затрагивается морфология части почечной системы), интеграцию графов знаний (например, автоматическая привязка к фенотипическим аномалиям из широко используемых онтологий, таких как онтология фенотипа человека (HPO) или фенотипа млекопитающих (MP)) и вывод знаний (например, логический вывод недостающих данных) (Dececchi et al., 2015).

Применение онтологий для интерпретации глубокого обучения

Глубокое обучение (DL) наглядно продемонстрировало свою эффективность при решении задач в области геномики, протеомики, биомедицины, включая анализ и автоматическую функциональную аннотацию последовательностей ДНК, РНК и белков, поиск ДНК/РНК-мишеней регуляторных РНК и белков, прогнозирование свойств и функций биомолекул, поиск 3D структуры белка, реконструкцию структур биомолекул с заданными свойствами, прогнозирование взаимодействия биомолекул и выявление на этой основе потенциальных кандидатов на лекарственные препараты, обработку и анализ изображений, интеграцию омиксных данных, анализ сложных, гетерогенных и взаимосвязанных биологических сетей (в том числе сетей белок-белковых взаимодействий, генных ре-

гуляторных сетей, семантических сетей и метаболических путей), моделирование биологических систем и процессов и т. д. (Li et al., 2019; Sapoval et al., 2022).

Одной из ключевых проблем глубокого обучения в биоинформатике, системной биологии и современной биомедицине является недостаточная интерпретируемость моделей нейронных сетей, которые часто функционируют как модели «черного ящика».

Интерпретируемость алгоритмов машинного обучения в биоинформатике и биомедицине важна по трем основным причинам. Во-первых, при анализе сложных систем, когда нет теории и четкого алгоритма принятия решений, требуется понять, почему модель делает такое предсказание. Во-вторых, важно гарантировать, что модель основывает свои прогнозы на надежном представлении данных и не фокусируется на нерелевантных артефактах. И наконец, модель с высокоточными предсказаниями, возможно, выявила интересные закономерности, которые биологи хотели бы изучить.

В формально-логическом смысле интерпретация – это отображение формальной конструкции на сущности и их отношения, которые она представляет. В этом смысле можно сказать, что кто-то понимает формальную конструкцию, если может соотнести ее с соответствующими сущностями и предложениями реального мира и рассуждать о последствиях. При этом важно отличать понятность модели от понимания того, почему модель истинна или как модель была получена из данных.

Можно выделить два основных подхода к интерпретации черных ящиков: апостериорные методы и самообъясняемые модели (Adadi, Berrada, 2018). В апостериорном методе сначала изучается модель «черного ящика», а затем используется метод интерпретации для объяснения того, как делались прогнозы. Часто полученные таким образом объяснения не совпадают с тем, как алгоритм глубокого обучения на самом деле получает решение. Кроме того, процедура объяснения является отдельным методом со своими ошибками, которые влияют на качество принимаемых решений. Поэтому такое объяснение не всегда подходит для биомедицины.

Необходимо отметить, что интерпретируемость – это понятие, специфичное для конкретной области, поэтому не может быть универсального определения. Очень часто в интерпретируемой модели машинного обучения добавляются ограничения в модельной форме, так чтобы она либо была полезна кому-то, либо подчинялась структурным знаниям области, таким как монотонность (Gupta et al., 2016), причинность, структурные (генеративные) ограничения, аддитивность (Lou et al., 2013) или физические ограничения, которые исходят из знания предметной области (онтологий).

К настоящему времени уже опубликовано несколько работ по построению самообъясняемых нейронных сетей на основе данных об экспрессии генов с использованием знания Gene Ontology. Так, например, в работе (Bourgeais et al., 2021) предложена самообъясняемая модель глубокого обучения Deep GONet, интегрирующая Gene Ontology в иерархическую архитектуру нейронной сети. Эта модель основана на полностью связанной архитектуре, ограниченной аннотациями GO, так что каждый нейрон представляет биологическую функцию. Эксперименты с наборами данных диагностики рака показывают, что Deep GONet легко интерпретируется и обладает высокой производительностью для различения раковых и нераковых образцов.

Другим примером самообъясняемой нейронной сети является ONN4MST – обобщение вычислительной модели Neural Network, основанной на онтологии (ONN), для отслеживания микробных источников (Zha, Ning, 2022). Модель ONN использует новый онтологический подход, который поощряет предсказание, удовлетворяющее онтологии «биомов». Другими словами, модель ONN может использовать информацию онтологии биома для моделирования зависимостей между биомами и оценки доли различных биомов в выборке сообщества.

Способность ONN4MST к открытию новых знаний продемонстрирована в различных приложениях отслеживания источников. Она позволяет отслеживать источники образцов, ниши которых ранее были менее изучены или неизвестны, обнаруживать микробные загрязнители, а также идентифицировать аналогичные образцы из онтологически удаленных биомов, показывающих уникальную важность ONN4MST в открытии знаний из огромного количества микробных образцов сообщества гетерогенных биомов.

ONN4MST может отличать образцы от онтологически близких биомов и, таким образом, предлагает количественный способ охарактеризовать развитие микробного сообщества кишечника человека. В частности, показано, что микробиом кишечника долгожителей отличается от микробиома нормальных пожилых людей и более похож на микробиом молодых людей (Zha, Ning, 2022).

Заключение

Бурное развитие экспериментальных технологий в области молекулярной биологии привело к тому, что онтологическое моделирование становится базовым методом в биоинформатике и системной биологии. Создание нескольких сотен базовых ссылочных онтологий и их верификация позволяют применять эти онтологии в качестве источников знаний для интеграции и построения сложных моделей предметной области и баз знаний, ориентированных на решение конкретных задач биомедицины.

Особое значение онтологии имеют для интерпретации результатов компьютерных предсказаний, полученных с помощью методов глубокого обучения. Чтобы ученые доверяли глубокому обучению, которое часто представлено как модели «черного ящика», должны использоваться специальные методы интерпретации на основе дополнительных знаний о предметной области или онтологий. Онтологии, паттерны их конструирования, интеграция больших данных и создание графов знаний играют ключевую роль в повышении интерпретируемости моделей глубокого обучения. Эти инструменты позволяют не только улучшить понимание результатов, но и обеспечить более качественный анализ данных. В условиях стремительного роста объемов информации и сложности моделей глубокого обучения использование онтологий становится необходимым шагом к созданию более прозрачных и объяснимых систем.

Можно ожидать, что новое поколение систем интерпретации будет не только в состоянии объяснить полученные решения понятным для человека способом, с указанием количественного уровня неопределенности, но и предложит дополнительные шаги (например, дополнительные эксперименты, клинические исследования и т. д.), необходимые для уточнения или надежного подтверждения своих решений.

Список литературы / References

Подколодный Н.Л., Игнатьева Е.В., Подколодная О.А., Колчанов Н.А. Информационная поддержка исследования механизмов регуляции транскрипции: онтологический подход. *Вавиловский журнал генетики и селекции*. 2012;16(4/1):742-755 [Podkolodnyy N.L., Ignatyeva E.V., Podkolodnaya O.A., Kolchanov N.A. Information support of research on transcriptional regulatory mechanisms: an ontological approach. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2012;16(4/1):742-755 (in Russian)]

Adadi A., Berrada M. Peeking inside the Black-Box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-52160. doi 10.1109/ACCESS.2018.2870052

Bergmann F.T., Czauderna T., Dogrusoz U., Rougny A., Drager A., Toure V., Mazein A., Blinov M.L., Luna A. Systems biology graphical notation markup language (SBGNML) version 0.3. *J. Integr. Bioinform*. 2020;17(2-3):20200016. doi 10.1515/jib-2020-0016

Bourgeais V., Zehraoui F., Ben Hamdoune M., Hanczar B. Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data. *BMC Bioinformatics*. 2021;22(S10):455. doi 10.1186/s12859-021-04370-7

Callahan T.J., Tripodi I.J., Stefanski A.L., Cappelletti L., Taneja S.B., Wyrwa J.M., Casiraghi E., Matentzoglu N.A., Reese J., Silverstein J.C., Hoyt C.T., Boyce R.D., Malec S.A., Unni D.R., Joachimiak M.P., Robinson P.N., Mungall C.J., Cavalleri E., Fontana T., Valentini G., Mesiti M., Gillenwater L.A., Santangelo B., Vasilevsky N.A., Hoehndorf R., Bennett T.D., Ryan P.B., Hripcsak G., Kahn M.G., Bada M., Baumgartner W.A., Hunter L.E. An open source knowledge graph ecosystem for the life sciences. *Sci. Data*. 2024;11(1):363. doi 10.1038/s41597-024-03171-w

Caufield J.H., Putman T., Schaper K., Unni D.R., Hegde H., Callahan T.J., Cappelletti L., Moxon S.A.T., Ravanmehr V., Carbon S., Chan L.E., Cortes K., Shefchek K.A., Elsarboukh G., Balhoff J., Fontana T., Matentzoglu N., Bruskiewich R.M., Thessen A.E., Harris N.L., Munoz-Torres M.C., Haendel M.A., Robinson P.N., Joachimiak M.P., Mungall C.J., Reese J.T. KG-Hub – building and exchanging biological knowledge graphs. *Bioinformatics*. 2023;39(7):btad418. doi 10.1093/bioinformatics/btad418

Chandrasekaran B., Josephson J., Benjamins V. What are ontologies, and why do we need them? *IEEE Intell. Syst. Appl*. 1999;14(1):20-26. doi 10.1109/5254.747902

Cooper L., Jaiswal P. The plant ontology: a tool for plant genomics. In Edwards D. (Ed.) Plant Bioinformatics. Methods in Molecular Biology. Vol. 1374. New York: Humana Press, 2016;89-114. doi 10.1007/978-1-4939-3167-5_5

Dececchi T.A., Balhoff J.P., Lapp H., Mabee P.M. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Syst. Biol*. 2015;64(6):936-952. doi 10.1093/sysbio/syv031

Diehl A.D., Meehan T.F., Bradford Y.M., Brush M.H., Dahdul W.M., Dougall D.S., He Y., Osumi-Sutherland D., Ruttenberg A., Sarntivijai S., Van Slyke C.E., Vasilevsky N.A., Haendel M.A., Blake J.A., Mungall C.J. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*. 2016;7(1):44. doi 10.1186/s13326-016-0088-7

Gkoutos G.V., Schofield P.N., Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. *Brief Bioinform*. 2018;19(5):1008-1021. doi 10.1093/bib/bbx035

Gupta M., Cotter A., Pfeifer J., Voevodski K., Canini K., Mangylov A., Moczydlowski W., van Esbroeck A. Monotonic calibrated interpolated look-up tables. *J. Mach. Learn. Res*. 2016;17:1-47

Hastings J., Owen G., Dekker A., Ennis M., Kale N., Muthukrishnan V., Turner S., Swainston N., Mendes P., Steinbeck C. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res*. 2016;44(D1):D1214-D1219. doi 10.1093/nar/gkv1031

Huntley R.P., Sawford T., Mutowo-Meullenet P., Shypitsyna A., Bonilla C., Martin M.J., O'Donovan C. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res*. 2015;43(D1):D1057-D1063. doi 10.1093/nar/gku1113

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol*. 2015;9(Suppl.2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. doi 10.1186/s12859-018-2567-6

Li Y., Huang C., Ding L., Li Z., Pan Y., Gao X. Deep learning in bioinformatics: introduction, application, and perspective in big data era. *Methods*. 2019;166:4-21. doi 10.1016/j.ymeth.2019.04.008

Livingston K.M., Bada M., Baumgartner W.A., Hunter L.E. KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*. 2015;16(1):126. doi 10.1186/s12859-015-0559-3

Lobentanzer S., Aloy P., Baumbach J., Bohar B., Carey V.J., Charoentong P., Danhauser K., Doğan T., Dreo J., Dunham I., Farr E., Fernandez-Torras A., Gyori B.M., Hartung M., Hoyt C.T., Klein C., Korcsmaros T., Maier A., Mann M., Ochoa D., Pareja-Lorente E., Popp F., Preusse M., Probul N., Schwikowski B., Sen B., Strauss M.T., Turei D., Ulusoy E., Waltemath D., Wodke J.A.H., Saez-Rodriguez J. Democratizing knowledge representation with BioCypher. *Nat. Biotechnol*. 2023;41(8):1056-1059. doi 10.1038/s41587-023-01848-y

Lou Y., Caruana R., Gehrke J., Hooker G. Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Assoc. for Computing Machinery, 2013;623-631. doi 10.1145/2487575.2487579

Mungall C.J., Torniai C., Gkoutos G.V., Lewis S.E., Haendel M.A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 2012;13(1):R5. doi 10.1186/gb-2012-13-1-r5

Osumi-Sutherland D., Courtot M., Balhoff J., Mungall C. Dead simple OWL design patterns. *J. Biomed. Semant*. 2017;8:18. doi 10.1186/s13326-017-0126-0

Podkolodnyy N.L., Podkolodnaya O.A. Ontologies in bioinformatics and systems biology. *Russ. J. Genet. Appl. Res.* 2016;6(7):749-758. doi 10.1134/S2079059716070091

Qaiser A., Ghulam S. Bioinformatics and big data analytics in genomic research. *Med. Pap.* 2023;3(1):165-179. doi 10.31219/osf.io/5grpc

Santos A., Colaço A.R., Nielsen A.B., Niu L., Strauss M., Geyer P.E., Coscia F., Albrechtsen N.J.W., Mundt F., Jensen L.J., Mann M. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol*. 2022;40(5):692-702. doi 10.1038/s41587-021-01145-6

Sapoval N., Aghazadeh A., Nute M.G., Antunes D.A., Balaji A., Baraniuk R., Barberan C.J., Dannenfelser R., Dun C., Edrisi M., Elworth R.A.L., Kille B., Kyrillidis A., Nakhleh L., Wolfe C.R., Yan Z., Yao V., Treangen T.J. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun*. 2022;13(1):1728. doi 10.1038/s41467-022-29268-7

Slater L.T., Gkoutos G.V., Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. *BMC Med. Inform. Decis. Mak*. 2020;20(Suppl.10):311. doi 10.1186/s12911-020-01336-2

Smith B., Ceusters W., Klagges B., Kohler J., Kumar A., Lomax J., Mungall C., Neuhaus F., Rector A.L., Rosse C. Relations in biomedical ontologies. *Genome Biol*. 2005;6(5):R46. doi 10.1186/gb-2005-6-5-r46

Stefancsik R., Balhoff J.P., Balk M.A., Ball R.L., Bello S.M., Caron A.R., Chesler E.J., de Souza V., Gehrke S., Haendel M., Harris L.W., Harris N.L., Ibrahim A., Koehler S., Matentzoglu N., McMurry J.A., Mungall C.J., Munoz-Torres M.C., Putman T., Robinson P., Smedley D., Sollis E., Thessen A.E., Vasilevsky N., Walton D.O., Osumi-Sutherland D. The Ontology of Biological Attributes (OBA)-computational traits for the life sciences. *Mamm. Genome*. 2023;34(3):364-378. doi 10.1007/s00335-023-09992-1

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E. Big Data: astronomical or genomical? *PLoS Biol*. 2015;13(7):e1002195. doi 10.1371/journal.pbio.1002195

Thomas P.D., Hill D.P., Mi H., Osumi-Sutherland D., Van Auken K., Carbon S., Balhoff J.P., Albou L.-P., Good B., Gaudet P., Lewis S.E., Mungall C.J. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet*. 2019;51(10):1429-1433. doi 10.1038/s41588-019-0500-1

Wood E.C., Glen A.K., Kvarfordt L.G., Womack F., Acevedo L., Yoon T.S., Ma C., Flores V., Sinha M., Chodpathumwan Y., Termehchy A., Roach J.C., Mendoza L., Hoffman A.S., Deutsch E.W., Koslicki D., Ramsey S.A. RTX-KG2: a system for building a semantically standardized knowledge graph for translational biomedicine. *BMC Bioinformatics*. 2022;23(1):400. doi 10.1186/s12859-022-04932-3

Zha Y., Ning K. Ontology-aware neural network: a general framework for pattern mining from microbiome data. *Brief. Bioinform*. 2022;23(2):bbac005. doi 10.1093/bib/bbac005

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 28.10.2024. После доработки 08.11.2024. Принята к публикации 11.11.2024.