ВЫСТУПЛЕНИЕ АКАДЕМИКА РАН Н.А. КОЛЧАНОВА

Материал поступил в редакцию 03.12.2018 г.
Принят к публикации 25.12.2018 г.

Ключевые слова: Большие данные, науки о жизни, искусственный интеллект, машинное обучение, нейронные сети, генные сети, когнитивные компьютерные системы.

DOI: https://doi.org/10.31857/S0869-5873894373-375

Стремительное развитие информационно-вычислительных технологий и их повсеместное внедрение практически во все сферы жизни стали ключевым фактором колоссального роста данных, который привёл к информационному взрыву, затронувшему буквально все области, включая науки о жизни. Так, согласно аналитическим исследованиям, уже к 2025 г. объём одних только геномных данных за счёт развития высокопроизводительных технологий секвенирования может в несколько раз превысить совокупный объём информации, продуцируемой астрономией и социальными сетями YouTube и Twitter. Повсеместное внедрение в клиническую практику электронных медицинских карт, развитие методов медицинской визуализации (рентгенография, МРТ, ультразвук и др.), а также прогресс в сфере диагностики состояния здоровья пациентов на основе приборов, оснащённых сенсорами, оказали существенное влияние на возникновение Больших данных в медицине. Развитие сенсорных и смарт-технологий, а также беспилотных аппаратов, снабжённых средствами фото/видеофиксации, способствовало появлению Больших данных в такой области, как сельское хозяйство. Кроме того, существуют десятки тысяч фактографических баз данных и десятки миллионов текстов патентов, содержащих ценнейшую информацию о живых системах. Только в базе данных PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) сегодня проиндексировано более 29 млн рефератов научных публикаций, посвящённых биологии и биомедицине, и это число продолжает неуклонно расти.

Стремительный рост, колоссальные объёмы, а также слабая структурированность и неоднородность подавляющей части подобной информации, нередко сочетающиеся с её зашумлённостью, делают применение лишь традиционных способов анализа данных недостаточно эффективным. Данная проблема стала поворотным моментом в развитии подходов машинного обучения и искусственного интеллекта, позволяющих автоматически выявлять скрытые взаимосвязи и закономерности в больших наборах данных, которые далеко не всегда очевидны для человека. Положительная динамика публикаций, связанных с применением методов искусственного интеллекта и машинного обучения в таких областях, как генетика, медицина, биотехнологии и сельское хозяйство, хорошо иллюстрирует увеличение интереса к этим технологиям со стороны научного сообщества.

Для генетики одним из бурно развивающихся направлений является создание систем, обеспечивающих реконструкцию генных сетей на основе информации, автоматически экстрагированной из фактографических баз данных и текстов научных публикаций. При этом под генными сетями понимаются группы координированных функционирующих генов, контролирующих формирование фенотипических характеристик организмов (молекулярных, биохимических, клеточных, физиологических, морфологических и др.). Неотъемлемые компоненты генных сетей – сети белок-белковых взаимодействий, метаболические пути, а также пути передачи сигналов. Реконструкция и анализ таких сетей имеют важнейшее значение для широкого ряда областей знаний и практических применений, включая биомедицину, фармакологию, биотехнологию, генетику, селекцию, сельское хозяйство и многие другие.

Один из хорошо известных примеров – когнитивная система Watson, разработанная компанией IBM и адаптированная для анализа информации из области наук о жизни. Основными задачами, на решение которых направлен этот инструмент, выступают экстракция и накопление знаний о биологических процессах и системах на основе применения методов автоматического анализа текстов и баз данных, реконструкция генных сетей, интерпретация биомедицинских данных, а также поиск фармакологических мишеней. База знаний системы содержит информацию о более чем 200 тыс. фенотипических признаков организмов, 21 тыс. химических соединений, 1300 лекарствах, 22 тыс. генов и сотнях тысяч белков. Модуль извлечения знаний из текстов научных публикаций и баз данных системы Watson реализован на основе применения ряда программ-аннотаторов, использующих нейронные сети. Такие сети обучены для выявления связей между объектами, относящимися к соответствующим предметным областям знаний.

Система Ingenuity Pathway Analysis (IPA), разработанная компанией QIAGEN, обеспечивает интеграцию, анализ и интерпретацию геномных, транскриптомных, протеомных, метаболомных

данных, поиск генов-мишеней и биомаркеров заболеваний. Её база знаний содержит около 5 млн фактов о генах, белках, метаболитах, лекарствах, а также взаимодействиях между ними, включая их связи с болезнями и биологическими функциями. Общая сеть взаимодействий IPA состоит из примерно 40 тыс. вершин и 1480 тыс. связей. Интеллектуальная компонента системы обеспечивает, во-первых, формирование базы знаний с использованием методов автоматического анализа научных публикаций и баз данных; во-вторых, автоматическую генерацию гипотез о генных сетях, путях передачи сигналов и метаболических путях, ответственных за изменения экспрессии генов, которые наблюдаются в экспериментах; в-третьих, выявление регуляторов верхнего уровня (миРНК, транскрипционные факторы, метаболиты, лекарства), вызывающих наблюдаемые изменения экспрессии генов.

Другими примерами хорошо известных систем являются STRING, Pathway Studio и MetaCore (https://clarivate.com/products/metacore/). Система STRING предназначена для реконструкции и анализа сетей белок–белковых взаимодействий, включая прямые (физические) и косвенные (функциональные) ассоциации. База знаний STRING содержит информацию о более чем 1380 млн взаимодействий для 9,6 млн белков из 2031 организма, полученную на основе экспериментальных данных, методов автоматического извлечения знаний из текстов научных публикаций, а также из курируемых баз.

Pathway Studio позволяет осуществлять построение и анализ биологических путей, сетей генной регуляции и сетей белок-белковых взаимодействий. Система включает данные о более чем 7 млн молекулярных взаимодействий, автоматически извлечённых из абстрактов и полных текстов научных публикаций из курируемых баз данных, а также данные о более 2000 курируемых вручную путей.

Система MetaCore обеспечивает интеграцию омиксных данных, комплексный анализ молекулярно-генетических сетей, быструю реконструкцию молекулярно-генетических и клеточных механизмов патогенеза, поиск биомаркеров и идентификацию лекарственных мишеней. В базе знаний этой системы содержится информация о более чем 1,7 млн межмолекулярных взаимодействий, 1600 биологических путях, а также 230 тыс. ассоциаций с заболеваниями.

Все вышеописанные системы разработаны за рубежом. Единственная российская когнитивная система ANDSystem, предназначенная для работы с генными сетями, – в Институте цитологии и генетики СО РАН. Система обеспечивает реконструкцию и анализ сетей молекулярно-генетических взаимодействий, интерпретацию экспериментальных данных, поиск новых фармакологических мишеней, а также генов – кандидатов для генотипирования. Она позволяет выявлять гены, которые вносят максимальный вклад в формирование целевых фенотипических (клинических) признаков, контролируемых генными сетями, и на этой основе предсказывать наиболее перспективные мишени для терапии заболеваний. База знаний ANDSystem содержит информацию приблизительно о 2 млн генов и белков, 46 тыс. заболеваний, 80 тыс. метаболитов, 90 тыс. биологических процессов, 4,5 тыс. микроРНК и 30 млн взаимодействий, автоматически экстрагированных из 24 млн документов PubMed и внешних баз данных. Кроме того, ANDSystem ориентирована на решение такой важной для фармакологии задачи, как выявление генов-мишеней, имеющих определяющее значение для функционирования генных сетей и контролируемых ими фенотипических признаков. Система обладает большой прогностической силой, подтверждённой в ряде экспериментальных работ. В частности, на основе анализа структуры генной сети, описывающей механизмы коморбидного состояния[1] астма/гипертония и реконструированной с помощью ANDSystem, были предсказаны гены, играющие ключевую роль в развитии патологии.

Экспериментальный анализ генетических полиморфизмов, выполненный для трёх предсказанных генов, определённых ANDSystem в качестве наилучших кандидатов для генотипирования, показал, что нуклеотидные замены в регуляторных районах этих генов с повышенной частотой встречаются в группе пациентов с коморбидным состоянием астмы и гипертонии. Применение ANDSystem для построения и анализа генных сетей, описывающих молекулярные механизмы патологических процессов при туберкулёзе, позволило выявить новые гены, отвечающие за восприимчивость человека к туберкулёзу.

В целом на сегодняшний день при разработке систем искусственного интеллекта, предназначенных для решения различных задач из области наук о жизни, применяется огромное количество методов и алгоритмов, таких как методы опорных векторов, решающих деревьев, логическая регрессия, байесовские модели и др. Очень перспективным оказалось применение методов глубокого машинного обучения и глубоких нейронных сетей, продемонстрировавших свою эффективность в решении широкого спектра задач из области живых систем. Особенность данного класса методов – обучение представлениям об объектах на основе анализа большого объёма данных, играющих роль примеров. Уникальным для глубокого обучения является автоматическое определение признаков и их иерар-

[1] Под коморбидным состоянием понимается наличие у пациентов одновременно двух заболеваний с частотой их совместной встречаемости выше, чем можно ожидать по случайным причинам. В настоящее время такие состояния активно изучаются.

хическая структуризация, при которой представление о более сложных признаках формируется за счёт комбинирования более простых. Например, в системе DeepChrome глубокое машинное обучение использовалось для предсказания экспрессии генов по паттернам модификации белков. Авторы системы deepTarget успешно применили рекуррентную нейронную сеть для предсказания мишеней микроРНК в матричной РНК. Ещё один вариант – использование трёхмерной свёрточной нейронной сети для оценки энергии взаимодействия "белок–лиганд".

Методы глубокого машинного обучения широко применяются и в сельском хозяйстве. Так, нейронные сети обеспечивают высокоточное автоматическое обнаружение вредителей и заболеваний томатов в режиме реального времени. При этом для обучения нейронной сети используется сравнительно небольшая выборка из 5 тыс. изображений листьев томатов, повреждённых фитофторой и вредителями, которые были сделаны при разных условиях (температура, время года, уровень влажности и т. д.).

Практическая простота применения и наличие развитого инструментария делают нейронные сети мощнейшим инструментом современного анализа данных. В то же время ключевой недостаток инструментов, разрабатываемых на основе глубокого машинного обучения, – исключительно низкий уровень прозрачности принимаемых решений, то есть их интерпретируемости в рамках устоявшейся терминологии соответствующих предметных областей. Преодоление этого недостатка – важнейший из вызовов, стоящих перед теорией и практикой искусственного интеллекта.

Один из перспективных подходов к повышению прозрачности решений предполагает интеграцию нейронных сетей с онтологиями предметных областей. В 2018 г. была опубликована работа, выполненная учёными из Стэнфордской школы медицины, в которой на основе массива больших геномных, транскриптомных и протеомных данных, характеризующих 5 млн линий дрожжей с нокаутами генов, и информации из Онтологии генов (Gene Ontology – GO) была построена нейронная сеть для предсказания влияния нокаутов дрожжевых генов на скорость роста дрожжевой культуры. GO – одна из самых больших онтологий, содержащих универсальное формализованное описание молекулярно-генетических функций, структур и процессов. В основе GO лежат три независимых раздела: биологические структуры (4202 сущности, 2 044 124 аннотации), молекулярные функции (11 150 сущностей, 2 001 539 аннотаций) и биологические процессы (29 691 сущность, 3 188 847 аннотаций). Онтология построена по принципу ориентированного ациклического графа: каждый термин связан с одним или несколькими другими терминами через различный тип отношений: "A is a B" – A является частным случаем B, "A part of B" – A является частью B, "B has part A" – B включает A, "A regulates B" – A регулирует B, "A positively regulates B" – A положительно регулирует B, "A negatively regulates B" – A отрицательно регулирует B, "A occurs in B" – A встречается при B. Моделирование влияния двойного нокаута в геноме дрожжей по генам *CYT1* и *COX7* на скорость роста дрожжевой культуры, реализованное за счёт применения данной нейронной сети, показало, что двойной нокаут приводит к нарушению функционирования эндоплазматического ретикулума, которое сопровождается появлением денатурированных белков, и, как следствие, к снижению скорости роста мутантной дрожжевой культуры. Такой подход позволяет не только оценивать влияние мутаций на скорость роста дрожжей, но и выявлять ключевые процессы, лежащие в основе повреждающих эффектов.

В заключение следует отметить, что быстрое накопление огромных объёмов сложно организованных гетерогенных и распределённых данных в области живых систем требует существенной интенсификации работ междисциплинарного характера по использованию методов искусственного интеллекта и машинного обучения. Необходимо создать междисциплинарную комплексную программу научных исследований, которая объединит как специалистов в области математики и информатики, так и учёных, непосредственно применяющих эти инструменты для решения связанных с исследованием живых систем научно-прикладных задач.

SPEECH OF THE RAS ACADEMICIAN N.A. KOLCHANOV

Received: 03.12.2018
Accepted: 25.12.2018

Keywords: Big Data, life sciences, artificial intelligence, computer-assisted learning, neural networks, gene networks, cognitive computer systems.