АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2:54-3

Н.Ю. Бизюкова, О.А. Тарасова, А.В. Рудик, Д.А. Филимонов, В.В. Поройков

Автоматическое распознавание названий химических соединений в текстах научных публикаций*

Рассмотрены методы поиска и извлечения наименований низкомолекулярных химических соединений и данных об их экспериментально подтверждённой биологической активности из текстов научных публикаций. Проанализированы разработанные и опубликованные в течение последних десяти лет подходы для автоматизированного извлечения химической и биологической информации, представленной (а) наименованиями химических соединений и (б) наименованиями белков, генов и ассоциированных с ними видов биологической активности. Такие данные могут быть применены для идентификации и хранения названий химических соединений, включая все их возможные синонимы. Тематика научных публикаций весьма разнообразна, поэтому извлеченные данные о названиях химических соединений могут быть применены для получения информации о (1) способах синтеза определённого химического соединения; (2) его физико-химических свойствах; (3) его взаимодействии с высокомолекулярными соединениями (белками, мРНК животных и человека, и пр.) или проявлении им определённого вида биологической активности; (4) его терапевтических свойствах и данных клинических исследований.

Ключевые слова: *интеллектуальный анализ текстов, наименования химических соединений, информационный поиск*

DOI: 10.36535/0548-0027-2020-11-5

ВВЕДЕНИЕ

Процесс извлечения данных из слабо структурированных и формализованных текстов научных публикаций требует немалых усилий и затрат времени, особенно в условиях работы с большим объемом информации. В связи с этим возникает необходимость интеллектуального анализа текстов с помощью машинных методов их автоматизированной обработки, которые могут применяться в различных областях, включая медицину и биологию. Одно из направлений, требующих работы с большими массивами структурированных данных, – это биоинформатика. Полученные в результате анализа данные могут быть применены, в частности, для оценки биологической активности и токсичности химических соединений, а это необходимо при разработке новых лекарственных препаратов.

Методы извлечения данных из текстов рассмотрены в нескольких публикациях, включая обзор по методам обработки текстов биомедицинской тематики [1], а также описание и критический анализ методов извлечения данных из статей биологической и медицинской направленности [2–5].

В настоящей статье мы проанализируем методы извлечения названий химических соединений (ХС) из текстов научных публикаций, разработанные и опубликованные в течение последних десяти лет и рассмотрим методы автоматического извлечения данных о взаимодействии ХС с белками человека, приводящие к конкретным биологическим эффектам. В отличие от ранее опубликованных работ нами подробно рассматриваются корпуса (коллекции текстов), специально подготовленные для применения методов извлечения данных, а также способы представления текстов для их обработки компьютерными методами и алгоритмы извлечения данных о названиях химических соединений.

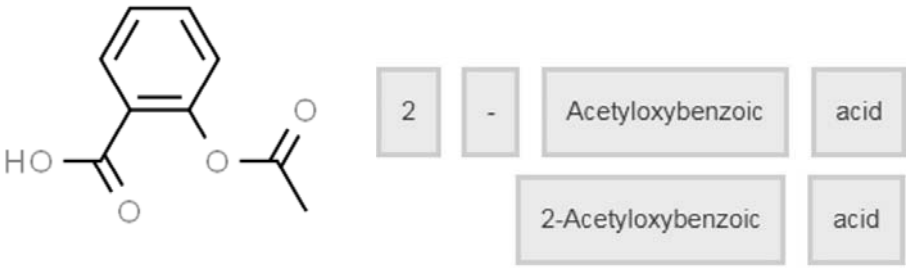
ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ ТЕКСТОВ

В интеллектуальном анализе текстов для извлечения данных о химических соединениях применяются как стандартные методы автоматизированной обработки текстов, так и специальные алгоритмы, направленные на поиск слов и словосочетаний, которые могут являться данными о ХС или их свойствах.

* Работа выполнена при поддержке гранта Российского научного фонда № 19-15-00396.

Рассмотрим представленные в электронном виде тексты, которые можно получить из библиографических баз данных (БД), например, БД Medline [6] в форматах PDF (Portable Document Format – межплатформенный открытый формат электронных документов, разработанный Adobe Systems Inc.), HTML (HyperText Markup Language – язык гипертекстовой разметки) и XML (eXtensible Markup Language — расширяемый язык разметки для автоматизированного создания и обработки документов).

Можно выделить следующие стандартные методы анализа текстов: предварительная обработка (препроцессинг), включающая (1) конвертацию наиболее распространённых форматов в простой текстовый формат; (2) разделение текста на элементарные единицы текста – токены, которые могут быть представлены как отдельными словами, так и различными символами и знаками препинания (токенизация); (3) классификацию токенов по принадлежности к частям речи (установку тегов соответствующих частей речи, Parts of Speech tags – PoS), приведение отдельных слов к словарной форме (лемматизация), удаление слов и терминов, которые часто встречаются, но не отражают смысловое содержание текстов узкоспециализированной, например, химической тематики (так называемые стоп-слова).



2-Acetyloxybenzoic acid

Рис. 1. Пример двух вариантов токенизации для ацетилсалициловой кислоты (наименование ИЮПАК 2-Acetyloxybenzoic acid).
Информация о вариантах наименований ацетилсалициловой кислоты получена из БД PubChem (https://pubchem.ncbi.nlm.nih.gov/).

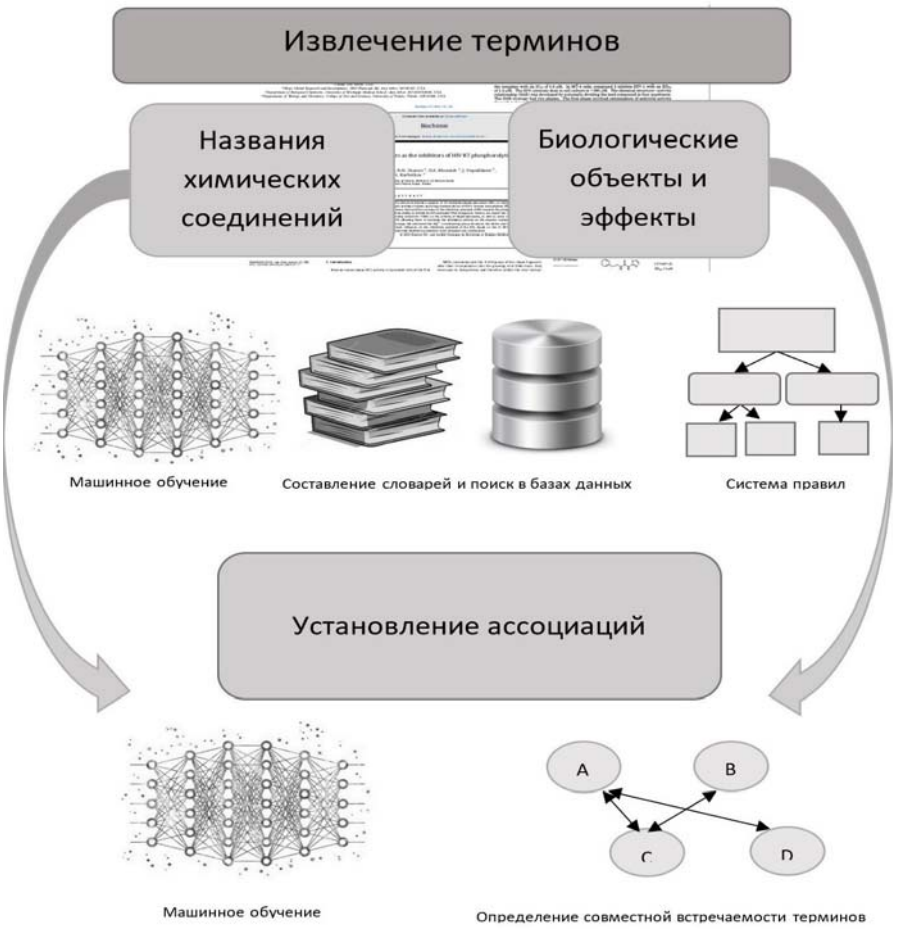


Рис. 2. Принципы работы алгоритмов по извлечению информации из текстов и установлению ассоциаций между терминами.

Токенизация – это разделение текста на слова, числа или другие специальные обозначения, или выделение нескольких слов, потенциально относящихся к одному и тому же термину. Алгоритм токенизации тривиален для простого текста с очевидными разделителями слов (пробелами). На практике проблемы токенизации возникают, когда в границах слова есть дефисы, запятые, скобки, апострофы – все это довольно типичная ситуация для названий химических соединений (рис. 1). Алгоритмы токенизации, как правило, базируются на применении методов машинного обучения к большому массиву текстов, где разметка токенов выполнена вручную. Выбор алгоритма токенизации важен при формировании словарей синонимов названий ХС на основе текстов научных публикаций [7, 8].

Разработку методов извлечения из текстов данных о химических соединениях и определения их взаимодействия с биологическими объектами можно условно разделить на две самостоятельные обширные задачи: (1) извлечение данных о ХС и биологических объектах и (2) поиск ассоциаций между ними (рис. 2).

ИЗВЛЕЧЕНИЕ ДАННЫХ О ХИМИЧЕСКИХ СОЕДИНЕНИЯХ И БИОЛОГИЧЕСКИХ ОБЪЕКТАХ

Распознавание в текстах научных публикаций наборов символов, обозначающих названия химические соединения, обычно относят к классу задач распознавания так называемых поименованных сущностей. Мы будем использовать общепринятые аббревиатуры: NER – Named Entity Recognition, NE – Named Entity, CNER – Chemical Named Entity Recognition (распознавание названий химических соединений) [1].

Рассмотрим основные источники данных для извлечения информации о названиях и свойствах химических соединений, являющихся, главным образом, коллекциями текстов, которые принято называть корпусами. Они свободно доступны для загрузки и анализа.

На основе больших БД библиографической информации [6, 9] исследователями создаются корпусы, которые содержат размеченные под определённые задачи тексты различных тематик. Тексты таких корпусов могут содержать метки соответствия каждого слова определённому термину, например, наименованию ХС, белку, гену, какому-либо биологическому эффекту, а также терминам, обозначающим взаимодействия между объектами в тексте. Так, корпус CHEMDNER разработан консорциумом научных групп из тридцати четырёх различных организаций и представляет собой библиотеку текстов из более 10 тыс. рефератов БД NCBI PubMed, содержащих (на 10.08.2020) 84 355 наименований химических соединений, которые вручную аннотированы экспертами. Отбор текстов производили таким образом, чтобы они отражали специфику основных направлений химии [7].

Для поиска ассоциаций между ХС и индуцируемыми ими заболеваниями был создан корпус CDR (chemical-disease relation extraction), содержащий свыше 1,5 тыс. аннотированных экспертами публикаций с данными о конкретных заболеваниях, названиях химических соединений и ассоциациях между ними [10, 11].

Авторами BEL (biological expression language) разработан корпус для поиска ассоциаций между ХС и белками, заболеваниями и биологическими процессами (звеньями патогенеза конкретных заболеваний). На 10.08.2020 этот корпус содержит 11 тыс. кратких описаний ассоциаций «белок–ХС» и «заболевание–ХС» из более чем 6 тыс. текстов [12].

Корпус DrugNer [13] содержит тексты с размеченными названиями торговых наименований лекарственных препаратов, а также аннотации различных видов биологических эффектов, которые могут вызывать эти препараты. Тексты корпуса представлены в формате XML, один текст – один файл XML, разделены на предложения и пронумерованы. Для каждого предложения выделен фрагмент (фраза), который размечен в соответствии с содержанием.

Корпусы с данными по взаимодействию лекарственных препаратов (drug-drug interactions – DDI) [14] (версии 2013 г.) помимо торговых наименований препаратов, содержат данные о межлекарственном взаимодействии с белками человека, в том числе с ферментами, которые участвуют в биотрансформации химических соединений.

Все приведенные здесь корпусы доступны для загрузки и обработки. Помимо этих корпусов создаются новые – для решения конкретных научных задач, стоящих перед исследователями, поэтому общее количество корпусов постоянно растёт. В табл. 1 представлены основные корпусы, разработанные в последнее десятилетие для решения задач извлечения из научных публикаций наименований химических соединений и поиска ассоциаций между ними, а также между биологическими объектами и процессами.

Способы хранения массивов текстов отобранных для конкретных задач различны. Но на практике, как правило, используются реляционные базы данных. Преимущество такого способа – возможность доступа к хранилищу сразу нескольким пользователям, вариативность форматов используемых таблиц в рамках различных СУБД, что влияет на скорость обработки запросов.

Методы распознавания названий химических соединений CNER включают: (1) методы на основе словарей и/или систем грамматических и лексических правил [1, 8, 18, 19] и (2) методы машинного обучения, разработанные или модифицированные для задач автоматизированного анализа текстов.

Для реализации многих методов, использующих системы правил, применяют упорядоченные множества терминов (словари), для которых известно, что они являются Named Entity. Сочетание поиска термина в словаре с правилами распознавания определённой последовательности слов и терминов (паттерна) позволяет выявить NE. Основное ограничение методов, основанных на системе правил – их сравнительно узкая область применимости, поскольку невозможно обнаружить в тексте наименования, которые отсутствуют в словаре (например, редко используемые синонимы или аббревиатуры, введенные авторами статьи) или не встречаются в определённой последовательности слов, для распознавания которой созданы правила.

Таблица 1

Источники данных для поиска ассоциаций между химическими соединениями, белками и патологическими процессами у человека

Название корпуса	Аннотированные типы биологических объектов и их взаимодействия	Количество документов	Дата последнего обновления	Ссылка
CEMP	Наименования ХС, генов, белков в текстах патентов	Свыше 1 000 наименований химических субстанций	2017	[15]
BEL	ХС, заболевания, патологические процессы	Более 11 000 коротких описаний из более 6 000 текстов	2016	[12]
BioCreative V Chemical Disease Relation (BC5 CDR) corpus	Болезни, белки/гены, ХС	1 500	2015	[10, 11]
CHEMDNER	ХС, белки	Свыше 10 000	2013	[7]
DDIExtraction	Тексты, содержащие названия лекарственных препаратов	Свыше 1 000 текстов	2013	[14]
CRAFT	Тексты, содержащие разметку наименований ХС	67 текстов	2012	[16]
ИЮПАК training corpus	ХС	~1 500 абстрактов	2008	[17]
DrugNer	Лекарственные соединения	885 абстрактов	2008	[16]

Класс задач об извлечении данных из текстов относится к интеллектуальному анализу (text mining), в котором широко используют методы искусственного интеллекта и методы машинного обучения, базирующиеся на представлении исследуемого текста в виде специфических признаков, отражающих символы, слова или фразы.

ПРЕДСТАВЛЕНИЕ СЛОВ ПРИ АНАЛИЗЕ ТЕКСТА

В методах классификации, базирующихся на машинном обучении, используется набор признаков, характеризующих текст. Для распознавания категории, содержания и контекста фраз обычно применяется набор свойств отдельных слов, включающий, например, метку принадлежности к определённой части речи, к численным значениям (а также последовательность буквенных символов и цифр), заглавным (прописным) буквам (ортографические признаки текста) или последовательности символов конкретных слов (так называемые *n*-граммы), а также содержащий особенности морфологии конкретных слов (например, наличие или отсутствие префиксов, выделение минимальной смысловой части слова и т.п. – морфологические признаки). В некоторых методах применяются метки наличия или отсутствия в исследуемом тексте фиксированных смысловых сочетаний слов, характерных для языка (табл. 2), а также их комбинации [1].

Один из распространённых видов представления слов в задачах NER – признаки вида BIO, где B (*beginning*) – метка начала поименованной сущности (NE), I (*inside*) – метка принадлежности слова к NE, O (*outside*) – метка, означающая, что соответствующее слово не относится к NE. Помимо указанных признаков (см. табл. 2) применяются разметки слов, в которых содержится информация о том, относится ли данный термин к названию ХС в тексте, а также метки принадлежности одного слова или нескольких слов (сочетаний слов) к названию химического соединения [1].

В ряде исследований [1, 8, 19] показано, что применение специфических типов признаков, разработанных для конкретной задачи, является предпочтительным по сравнению с использованием сочетания признаков, которыми обычно представляют текст любой тематики.

Ещё один вариант признаков, применяемых при анализе текстов, – это векторное представление слов (word embeddings), в котором используется приведение каждого слова и отдельных фраз, последовательно генерируемых из предложения, в бинарный вектор-строку. Способы генерации векторов могут быть разнообразными, например, они могут отражать частоту слов в исследуемых корпусах. При этом подходе контекст, в котором находится конкретное слово, может быть учтён посредством перевода в векторы нескольких слов, расположенных в непосредственной близости с этим словом, и определения общего вектора, вычисленного посредством применения какой-либо функции (в простейшем случае - усреднением). Задача сравнения контекстов фраз в этом случае сводится к сравнению двух векторов, отражающих контекстное представление фраз.

Помимо признаков, отражающих морфологию слова или принадлежность его к части речи, иногда могут использоваться значения, являющиеся результатами классификации после применения иных инструментов, т. е. не определённые экспертом, а расчётные значения [1].

Таблица 2

Признаки, применяемые для извлечения названий химических соединений

Признаки	Описание	Пример	Ссылки
Морфологические	Отражают строение слова (суффиксы, префиксы и т.п.)	2-(Acetyloxy)benzoic acid = Acetyl + oxy + benz + oic + acid	[20-23]
Лемма	Нормализованная форма слова (в именительном падеже единственного числа)	Drugs -> drug	[24-25]
POS	Принадлежность к частям речи	2-(Acetyloxy)benzoic acid = noun	[20-25]
Ортографические	Первая буква слова соответствует её виду в предложении (заглавная или прописная), признаки отображают количество символов, относящихся к буквам, цифрам и знакам препинания в рассматриваемом термине	2-(Acetyloxy)benzoic acid = 2 (первый символ слова)+ 20 (количество букв) + 1 (количество цифр)	[20-23]
«форма слова»	Буквы в слове представлены: заглавная буква - A, строчная буква – a, цифры – O, все остальные символы - o	2-(Acetyloxy)benzoic acid = OooAaaaaaaaaoaaaaaaaoaaaa	[20-23]
BIO	Метка принадлежности слова B - началу наименования ХС, I - термину в составе наименования ХС, O - термину, не относящемуся к ХС вообще	2-(Acetyloxy)benzoic acid is aspirin = BI O B	[20-23]
NO/NE/S-NE/M-NE/E	Метки, соответствующие: NO – слову, не являющемуся названием ХС NE – начало наименования ХС; S/M – слово/множество слов являющееся названием ХС E – окончание термина, соответствующего названию ХС	2-(Acetyloxy)benzoic acid is aspirin = NE + ME + E-NE + NO + SE	[20-23]

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТЕКСТОВ

На использовании значительного количества примеров, для которых есть данные об их принадлежности к конкретной категории основаны методы машинного обучения. Примеры представляются в виде совокупности признаков (описание и примеры признаков приведены выше). Таким образом, по сравнению с методами, основанными на системе правил, методы машинного обучения обеспечивают более высокую степень абстракции и обладают более широкой областью применимости. К основным применяемым при анализе текстов методам машинного обучения можно отнести методы, основанные на теореме Байеса, методы опорных векторов, методы, основанные на построении искусственных нейронных сетей (ИНС). Одним из наиболее часто применяемых в анализе текстов и поиске взаимосвязей между терминами, извлечёнными в результате анализа, является метод Conditional Random Fields (CRF) [26]. Методы, которые являются его вариациями, основаны на оценке контекста, т. е., в формальном приближении, последовательности слов. Метод CRF основан на представлении текста в виде ненаправленного графа, где последовательности слов являются вершинами, соединёнными связями. Целевой переменной может быть принадлежность фрагмента текста к конкретной категории, либо принадлежность конкретного слова или словосочетания к определённому типу терминов (например, названию ХС). Для каждого полного подграфа этого целенаправленного графа определяется потенциальная функция, которая каждому возможному состоянию элементов подграфа ставит в соответствие вещественное число. Кроме признаков каждого слова при оценке целевой переменной во внимание принимается значение целевой переменной на предшествующем этапе оценки текущего значения потенциальной функции. Таким образом в модели может быть учтён контекст.

Методы, основанные на применении искусственных нейронных сетей для NER, могут базироваться на стандартной архитектуре или различных модифицированных вариантах искусственных нейронных се-

тей. Каждый искусственный нейрон в ИНС представлен нелинейной функцией от комбинации входных сигналов. Соответственно, каждый нейрон может иметь несколько «входов» и один «выход». В ИНС стандартной архитектуры содержится один слой для входных сигналов (параметров, input layer), один или несколько так называемых «скрытых» слоев и один слой, в котором формируются результаты рассчитанных значений целевой переменной. Подробно теоретические основы и применение ИНС (для задач «структура-свойство» ХС) рассмотрены в обзорах [27, 28]. В NER используются модифицированные варианты НС [29], например, свёрточные ИНС (convolutional neural networks, CNN), рекуррентные ИНС [30], а также сети «долгой-краткосрочной памяти» (Long-Short-Term Memory, LSTM) [31]. Идея метода основана на том, что процесс мышления тесно связан с процессами памяти. Поэтому ИНС необходимо хранить информацию, полученную в процессе обработки входного сигнала, и при некоторых условиях трансформировать эту информацию. Особенностью архитектуры этой ИНС является наличие в ней переменной, сохраняющей состояния сети, рассчитанные на предшествующем слое. Такой подход позволяет учитывать контекст при NER (например, учитывать род прилагательного, находящегося в непосредственной близости с NE).

Методы NER на основе CRF могут применяться также в комбинации с другими способами сравнения текстовых выражений, такими, как например, метод нечёткого поиска или метод на основе ИНС [32].

Результаты классификации оцениваются стандартными метриками: полнотой (*sensitivity*, *recall*), точностью – доля истинно-положительных относительно всех признанных положительными результатов (*precision*, *positive predictive value*) и величиной *F-score* (F_1-*score*) – средним гармоническим *precision* и *recall*.

В проблеме анализа текстов для извлечения данных о конкретных свойствах химических соединений, например, ассоциациях между ХС, белками или видами биологической активности, можно выделить две самостоятельные задачи: (1) извлечение данных о наименованиях ХС, белков, генов или биологических процессов и (2) автоматическое установление ассоциаций между найденными терминами. Далее последовательно рассмотрим примеры алгоритмов для решения каждой из этих задач.

ИЗВЛЕЧЕНИЕ ДАННЫХ О ХИМИЧЕСКИХ СОЕДИНЕНИЯХ ПРИ АНАЛИЗЕ ТЕКСТОВ

Методы, основанные на искусственных нейронных сетях различной архитектуры, CRF, а также на комбинации методов машинного обучения описаны в работах [22, 32–36].

Авторы [33] применяли для анализа текстов метод CRF с последующим пост-процессингом. Ими разработан список правил, который позволяет осуществлять отбор наиболее вероятных истинно-положительных результатов. Множество правил основано на совпадении числа открывающихся и закрывающихся скобок в извлечённой последовательности слов, которая отнесена алгоритмом к названию химического соединения. Точность (*precision*) распознавания названий ХС составила 83,71%. Следует отметить, что, хотя пост-процессинг выбранных алгоритмом последовательностей слов позволяет снизить число ложноположительных результатов и таким образом увеличить точность NER, тем не менее, он не позволяет снизить число ложноотрицательных результатов. Фильтрация терминов посредством применения системы правил может быть осуществлена как после получения результатов работы ИНС, так и на первом этапе (при подготовке данных) исследования [1, 23, 32].

Авторы работы [32] применили метод CRF в комбинации с алгоритмом CNN-BiLSTM. Отличительная особенность их подхода в том, что работа свёрточных ИНС основана на признаках последовательности пяти символов текста, в то время как на основе BiLSTM анализируется большее число символов; соответственно, может быть проанализирован контекст целых фраз и предложений. Окончательная оценка принадлежности нескольких слов к Named Entity производится с применением метода CRF, на вход которого подаются значения, полученные с использованием CNN-BiLSTM. При этом в цитируемом подходе распознавались не только NE химических соединений, но и названия белков, организмов, клеточных линий и тканей. В этой же работе отмечаются, что распознавание одновременно нескольких типов NE (например, одновременно ХС, белок, организм, тканевая принадлежность, клеточная линия) существенно не влияет на оценки точности для тестовой выборки в сравнении с распознаванием какого-либо одного типа NE. Так, значение *precision* 0,775; *recall* 0,587; F_1-*score* 0,668 было получено при CNER в случае, если модель была обучена для распознавания одновременно нескольких NE, и соответствующие значения точности составили *precision* 0,661; *recall* 0,681; F_1-*score* 0,671 при CNER без распознавания других терминов.

В работе [34] реализована многоуровневая искусственная нейронная сеть, в которой на вход подаются векторные представления слов и отдельных символов. Авторы использовали три типа архитектур ИНС: свёрточную, рекуррентную и распределённую. Результатом классификации стали данные о принадлежности последовательности символов (слова) конкретному классу из наименований ХС (например, тривиальное название (TRIVIAL), систематическая номенклатура (SYSTEMATIC) и т.п.). Сочетание трёх типов ИНС позволило извлечь термины, относящиеся к названию химического соединения, и определить их класс. Точность классификации на тестовой выборке составила: *precision* 0,886, *recal* 0,888, F_1-*score* 0,887.

Широкое развитие методов машинного обучения позволяет извлекать из текста названия химических соединений, биологических объектов, процессов, а также определять ассоциации между ними. В то же время, поскольку при работе алгоритмов машинного обучения возможны ошибки ложно распознаваемых объектов [1], исследователи вынуждены вносить ограничения на работу алгоритма посредством построения системы правил, ранжирующих результаты

классификации, полученные методами машинного обучения. Таким образом, возникают комбинированные (или гибридные) подходы, которые основаны на методах машинного обучения вместе с применением совокупности правил или с осуществлением поиска в словарях терминов ХС.

В работе [35] авторы использовали модифицированный алгоритм BiLSTM-CRF, в котором совокупность признаков для отдельных слов подвергалась нескольким последовательным преобразованиям на каждом из этапов которых последовательно применяли функцию гиперболического тангенса, алгоритм CRF для каждого слова из предложения, причём на вход алгоритма CRF подавали последовательность из нескольких слов, которая была получена (1) при движении от начала к одному выбранному слову в предложении и (2) от конца к этому же слову, рассчитанную с учётом слов в предложении, предшествующих исследуемому. На конечном этапе каждое слово подавали на вход в алгоритм CRF, и затем оценивали результаты классификации. Максимальные параметры точности CNER для корпуса CHEMNDER при этом подходе составили: *precision* 0,917, *recall* 0,904.

Аналогично комбинированный подход реализован в методе LSTMVoter [36]. В LSTMVoter использован метод CRF в сочетании с искусственной нейронной сетью, основанной на архитектуре LSTM. На вход, помимо стандартных параметров, также подавались оценки классификации, рассчитанные другими методами CNER, разработанными ранее (табл. 3).

Среди комбинированных систем CNER также можно отметить ChemSpot [22]. В этом алгоритме метод CNER на основе CRF сочетали с поиском терминов, найденных в словаре Международного союза теоретической и прикладной химии IUPAC (International Union of Pure and Applied Chemistry). Авторы ChemSpot подчёркивают, что такой способ дает возможность однозначно определять соответствие найденному термину в словаре ИЮПАК, что, в свою очередь, позволяет избегать ошибок, опечаток, неточного названия в тексте при распознавании и сохранении термина ИЮПАК. Такой подход сочетает в себе элементы машинного обучения и метода, основанного на системе правил.

Подходы, используемые для поиска названий химических соединений, во многих случаях основаны на применении методов машинного обучения и искусственных нейронных сетей для интеграции результатов классификации, в том числе, полученных другими методами. Учитывая, что задача CNER в текстах научных публикаций является достаточно трудно формализуемой, машинное обучение может применяться совместно с подходами, основанными на системе словарей или системе правил фильтрации терминов, не имеющих отношения к названиям химических соединений (пост-процессинг).

Таблица 3

Примеры методов, реализующих поиск названий химических соединений в текстах научных публикаций

Цель метода	Корпус	Представление слов	Метод обучения	Средняя точность распознавания	Ссылка
CNER	CHEMNDER	Группировка символов на последовательность букв и цифр	Искусственные нейронные сети	Около 0,89	[34]
CNER	CHEMNDER CEMP	Разметка BIO	Комбинированный метод (LSTM-ANN-CRF)	0,89	[36]
NER ХС в совокупности с другими NE (белки, организмы и т.п.)	BIOCREATIVE (BIO-ID)	Разметка BIO, токенизация, стандартизация синонимов	Комбинированный метод (CNN-LSTM-CRF)	0,67	[32]
CNER	CHEMNDER	Комбинация признаков: - векторные представления - *n*-граммы - метки частей речи (POS tags)	Комбинированный метод (CRF)	0,90	[35]
CNER	MEDLINE	Последовательность из нескольких символов (n-граммы), коллокации (совместная встречаемость терминов), POS-tags	Conditional random fields	0,87	[33]
CNER	MEDLINE	Разметка слов по принадлежности к терминам ХС	Conditional random fields + поиск в словаре ИЮПАК	0,89	[22]

ПОИСК АССОЦИАЦИЙ МЕЖДУ ИЗВЛЕЧЁННЫМИ НАЗВАНИЯМИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ, БЕЛКАМИ И БИОЛОГИЧЕСКИМИ ПРОЦЕССАМИ

Автоматическое распознавание CNER может быть первым этапом в экстракции данных о взаимосвязи химических соединений с биологическими эффектами, которые они могут потенциально проявлять при условии, что автоматический поиск проводят в научных публикациях, в которых содержится экспериментальное подтверждение взаимодействия ХС с белками-мишенями или наличия определённого вида биологической активности. Разработка автоматических методов экстракции данных о ХС и их биологических эффектах или других свойствах может способствовать созданию обучающих выборок для методов анализа взаимосвязи «структура – активность» (Structure – Activity Relationship, SAR). Данные об экспериментально подтверждённых видах биологической активности химических соединений могут быть применены для поиска дополнительных фармакотерапевтических свойств лекарственных препаратов (Drug Repurposing).

Выявление различных биологических эффектов и/или взаимодействий между ХС и белками в литературе реализовано несколькими методами: (1) поиск названия активности/имени белка; (2) система правил, где указаны основные термины и характеристика взаимодействий белок-лиганд, описываемая с помощью этих терминов; (3) поиск на основе правил определённых «паттернов» – закодированных в определённой последовательности обобщённых терминов; (4) классификация фраз из текста на относящиеся или не относящиеся к описанию взаимодействий ХС с белками на основе методов машинного обучения; (5) определение совместной частоты встречаемости двух или нескольких терминов (например, названия лекарственного препарата и определённой биологической активности) в тексте. Далее приведем примеры установления ассоциаций между химическими соединениями и их биологическими свойствами.

В исследовании [37] предложения классифицировали на две категории: (1) содержащие описание взаимодействий между ХС и белком и (2) не содержащие такого описания. Обучающая выборка состояла из 1 632 аннотаций статей из БД научных публикаций NCBI PubMed. Предложения статьи обучающей выборки были проаннотированы и выбраны фразы, в которых содержались упоминания белков и химических соединений (пары NE), и фразы, в которых, дополнительно к парам NE, содержались упоминания о типе взаимодействий между белком и ХС. В качестве признаков использовали принадлежность аннотированных фраз к одному из семи типов взаимного расположения слов, описывающих термины (ХС, белок) в тексте, расстояние между терминами, выраженное в количестве слов, количество слов в предложении, наличие или отсутствие между терминами других слов биомедицинской тематики, признанных авторами подхода «значимыми» для выявления пар и триплетов NE, описывающих взаимодействие между ХС и белками. Классификация предложений основана на применении нескольких методов машинного обучения, включая деревья решений, наивный Байесов подход, логистическую регрессию, линейный дискриминантный анализ. Для определения принадлежности предложения к категории (1) или (2) было выявлено пересечение результатов отдельных классификаторов. Авторы [37] отмечают, что значения точности, полученные при реализации указанного метода (Precision 0,63; Recall: 0,51), сопоставимы с результатами, основанными на глубоком обучении искусственных нейронных сетей.

К настоящему времени разработаны и реализованы в виде веб-сервисов алгоритмы по выявлению потенциальных биологических эффектов для низкомолекулярных химических соединений и лекарственных препаратов. Большинство из этих алгоритмов нацелено на выявление одной или нескольких типов биологической активности, которые могут быть взаимосвязаны, например, токсичность и воздействие на ферменты метаболизма. Алгоритм LimTox (Literature Mining for Toxicology), реализованный в виде веб-сервиса), направлен на поиск биологических эффектов лекарственных препаратов, ассоциированных с гепатотоксичностью и взаимодействием с ферментами семейства цитохромов Р450. Поиск можно проводить и для других видов токсичности, таких как нефротоксичность, гепатотоксичность, кардиотоксичность [18].

В алгоритме, реализованном в ChemoText [38], существует возможность поиска терминов MeSH (Medical Subject Headings – это тезаурус словаря, контролируемый NLM (National Library of Medicine), используемый для индексации статей для PubMed) [39], ассоциированных с тремя типами NE: (1) низкомолекулярное химическое соединение; (2) белок; (3) заболевание. В ChemoText существует возможность поиска по терминам MeSH, анализа совместной встречаемости терминов, задаваемых пользователем, в списке терминов MeSH, а также построения с применением терминов MeSH семантических карт. Результаты поиска могут быть применены для (1) отбора публикаций, в которых исследуются конкретные низкомолекулярные ХС либо интересующий исследователя белок-мишень, и (2) для оценки ассоциаций между конкретными белками и заболеваниями и поиска препаратов, потенциально применимых для терапии этих заболеваний. Такой «двунаправленный» поиск и возможность отбирать публикации, в которых рассматриваются возможные звенья патогенеза заболеваний, а также химические соединения, предположительно ассоциированные с терапией заболевания, является преимуществом метода. В качестве ограничений можно отметить, что поиск производится только по MeSH терминам, исключая информацию из полного текста статьи. Кроме того, что при таком подходе теряется часть информации, известно, что термины MeSH доступны не для всех публикаций (они отсутствуют примерно в 10 % всех публикаций) в БД PubMed, что может приводить к ограничению множества анализируемых статей.

Подходы, объединяющие данные о химических соединениях и их возможных биологических эффектах, включающих взаимодействие с белками-мишенями, могут быть основаны и на сопоставлении названий белков-мишеней, найденных в тек-

стах публикаций, с таковыми из БД, содержащих информацию о них.

В работе Е. А. Пономаренко и соавторов были построены семантические сети для множества белков, ассоциированных с метаболическими путями [40]. Для построения семантических сетей авторы применяли выборку публикаций, которые были найдены в PubMed при добавлении названия белка в строку запроса (запросы сгенерированы автоматически). При выполнении поискового запроса были выявлены релевантные публикации для каждого белка. Затем определено пересечение множеств релевантных публикаций для пар исследуемых белков. На основании количества публикаций, релевантных для каждой пары исследуемых белков одновременно, и суммарного количества публикаций, релевантных хотя бы одному белку, были рассчитаны коэффициенты подобия (коэффициент Танимото). На основании функции, зависящей от значений коэффициента Танимото, были сформированы группы белков. Впоследствии были показаны общие функции для белков, входящих в состав одной группы, согласно предложенному методу. Разработанный алгоритм может быть применён и для выявления схожих свойств химических соединений.

ЗАКЛЮЧЕНИЕ

Автоматическое распознавание названий химических соединений в текстах научных публикаций позволяет решать ряд задач: автоматизированное пополнение баз данных о схемах химического синтеза, синтетической доступности ХС, в том числе, с требуемым видом биологической активности; поиск новых видов фармакологической активности для зарегистрированных лекарственных препаратов, а также возможных побочных эффектов, оценку вероятных межлекарственных взаимодействий.

Вместе с тем, CNER является достаточно трудоёмкой, сложно формализуемой задачей, поскольку тексты на любом языке имеют вариабельную структуру. Поэтому применяется ряд признаков, описывающих как непосредственно отдельные фрагменты текста (слова, символы), так и разметку, позволяющую учесть контекст. Помимо специфических признаков, контекст обычно учитывается при построении модели (например, в методе CRF возможно учитывать контекст; существуют варианты искусственных нейронных сетей, в которых предшествующее состояние сохраняется при прохождении по этим сетям). Точность распознавания (*precision*) названий химических соединений для большинства разрабатываемых методов извлечения наименований ХС находится в диапазоне 0,67 – 0,90. Точность CNER и других терминов биомедицинской тематики может быть существенно повышена, если в научных публикациях указывать принадлежность биологического или химического термина к классификации Bioassay Ontology [41] или Chemical Information Ontology [42]. Введение ограничений на обязательный конкретный формат названия химического соединения в статьях, например, формат ИЮПАК, позволяет увеличивать точность распознавания для задач NE. Помимо этого, разработка новых признаков для представления слов и применение дополнительных методов отбора релевантных публикаций для класса задач CNER может способствовать улучшению точности распознавания ассоциаций между названиями ХС и их вероятной биологической активностью, и, как следствие, увеличению объёма и повышению качества данных о свойствах химических соединений

СПИСОК ЛИТЕРАТУРЫ

1. Krallinger M., Rabal O., Lourenço A., Oyarzabal J., Valencia A. Information Retrieval and Text Mining Technologies for Chemistry // Chemical Reviews. – 2017. – Vol. 117, № 12. – P. 7673–7761.
2. Przybyła P., Shardlow M., Aubin S., Bossy R., Eckart de Castilho R., Piperidis S., McNaught J., Ananiadou S. Text mining resources for the life sciences // Database. – 2016. – Vol. 2016 (baw145), P. 1-30.
3. Oellrich A., Gkoutos G.V., Hoehndorf R., Rebholz-Schuhmann D. Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology // Journal of Biomedical Semantics. – 2012. – Vol. 3, № S2/S1. – P. 1-10.
4. O'Mara-Eves A., Thomas J., McNaught J., Miwa M., Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches // Systematic Reviews. – 2015. – Vol. 4, № 5. – P. 1-22.
5. Smink W.A.C., Fox J.-P., Tjong Kim Sang E., Sools A.M., Westerhof G.J., Veldkamp B.P. Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining // Frontiers in Psychology. – 2019. – Vol. 10. – P. 1186.
6. PubMed. – URL: https://pubmed.ncbi.nlm.nih.gov/
7. Krallinger M., Rabal O., Leitner F., Vazquez M., Salgado D., Lu Zh., Leaman R., Lu Y., Ji D., Lowe D.M., Sayle R. A., Batista-Navarro R.Th., Rak R., Huber T., Rocktäschel T., Matos S., Campos D., Tang B., Xu H., Munkhdalai T., Ryu K.H., Ramanan S.V., Nathan S., Žitnik S., Bajec M., Weber L., Irmer M., Akhondi S.A., Kors J.A., Xu Sh., An X., Sikdar K.U., Ekbal A., Yoshioka M., Dieb Th.M., Choi M., Verspoor K., Khabsa M., Giles C.L., Liu, H., Komandur Ravikumar K.E., Lamurias A., Couto F.M., Dai H.-D., Tzong-Han Tsai R., Ata C., Can T., Usié A., Alves R., Segura-Bedmar I., Martínez P., Oyarzabal J., Valencia A. The CHEMDNER corpus of chemicals and drugs and its annotation principles // Journal of Cheminformatics. – 2015. – Vol. 7, № S2. – P. 2-17.
8. Akhondi S.A., Hettne K.M., van der Horst E., van Mulligen E.M., Kors J.A. Recognition of chemical entities: combining dictionary-based and grammar-based approaches // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S6) – P. 1-10.
9. NCBI. – URL: https://www.ncbi.nlm.nih.gov/mesh/

10. Li J., Sun Y., Johnson R.J., Sciaky D., Wei C.-H., Leaman R., Davis A.P., Mattingly C.J., Wiegers T.C., Lu Z. BioCreative V CDR task corpus: a resource for chemical disease relation extraction // Database. – 2016. – Vol. 2016 (baw086). – P. 1-10.
11. Wei C.-H., Peng Y., Leaman R., Davis A.P., Mattingly C.J., Li J., Wiegers T.C., Lu Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task // Database. – 2016. – Vol. 2016 (baw032). P. 1-8.
12. Madan S., Szostak J., Komandur Elayavilli R., Tsai R.T.-H., Ali M., Qian L., Rastegar-Mojarad M., Hoeng J., Fluck J. The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2019) BEL track // Database. – 2019. – Vol. 2019 (baz084). – P. 1-17.
13. Martínez V., Navarro C., Cano C., Fajardo W., Blanco A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data // Artificial Intelligence in Medicine. – 2015. – Vol. 63, № 1. – P. 41-49.
14. Herrero-Zazo M., Segura-Bedmar I., Martínez P., Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions // Journal of Biomedical Informatics. – 2013. – Vol. 46, № 5. – P. 914-920.
15. Pérez-Pérez M., Rabal O., Pérez-Rodríguez G., Vazquez M., Fdez-Riverola F., Oyarzabal J., Valencia A., Lourenço A., Krallinger M. Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks // Proceedings of the BioCreative. Vers. 5. Challenge Evaluation Workshop. – 2017. – P. 11-18. – URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2017/ BioCreative_V5_paper2.pdf
16. Bada M., Eckert M., Evans D., Garcia K., Shipley K., Sitnikov D., Baumgartner Jr.W.A., Cohen B., Verspoor K., Blake J.A., Hunter L.E. Concept annotation in the CRAFT corpus // BMC Bioinformatics. – 2012. – Vol. 13, № 161. – P. 1-10.
17. Kola´rik C., Klinger R., Friedrich C.M., Hofmann-Apitius M., Fluck J. Chemical Names: Terminological Resources and Corpora Annotation // Workshop on Building and Evaluating Resources for Biomedical Text Mining (6th edition of the Language Resources and Evaluation Conference). – Marrakech (Morocco), 2008. – P. 51-58. – URL: https://pub.uni-bielefeld.de/record/2603498
18. Cañada A., Capella-Gutierrez S., Rabal O., Oyarzabal J., Valencia A., Krallinger M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes // Nucleic Acids Research. – 2017. – Vol. 45, № W1. – P. W484-W489.
19. Swain M.C., Cole J.M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature // Journal of Chemical Information and Modeling. – 2016. – Vol. 56, № 10.– P. 1894-1904.
20. Batista-Navarro R., Rak R., Ananiadou S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S6). – P. 1-13.
21. Leaman R., Khare R., Lu Z. Challenges in clinical natural language processing for automated disorder normalization // Journal of Biomedical Informatics. – 2015. – Vol. 57.–- P. 28-37.
22. Rocktäschel T., Weidlich M., Leser U. ChemSpot: a hybrid system for chemical named entity recognition // Bioinformatics. – 2012. – Vol. 28, № 12. – P. 1633-1640.
23. Campos D., Bui Q.-C., Matos S., Oliveira J.L. TrigNER: automatically optimized biomedical event trigger recognition on scientific documents // Source Code for Biology and Medicine. – 2014. – Vol. 9, №1. – P. 1.
24. Lu Z., Hirschman L. Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II // Database. – 2012. – Vol. 2012 (bas043). – P. 1-6.
25. Liu H., Christiansen T., Baumgartner W.A., Verspoor K. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text // Journal of Biomedical Semantics. – 2012. – Vol. 3, №3. – P. 1-29.
26. Song H.-J., Jo B.-C., Park C.-Y., Kim J.-D., Kim Y.-S. Comparison of named entity recognition methodologies in biomedical documents // BioMedical Engineering OnLine. – 2018. – Vol. 17 (Suppl 2). – P. 158-192.
27. Halberstam N.M., Baskin I.I., Palyulin V.A., Zefirov N.S. Neural networks as a method for elucidating structure–property relationships for organic compounds // Russian Chemical Reviews. – 2003. – Vol. 72, № 7. – P. 629-649.
28. Baskin I.I., Madzhidov T.I., Antipin I.S., Varnek A.A. Artificial intelligence in synthetic chemistry: achievements and prospects // Russian Chemical Reviews. – 2017. – Vol. 86, №11. – P. 1127-1156.
29. Cho H., Lee H. Biomedical named entity recognition using deep neural networks with contextual information // BMC bioinformatics. – 2019. – Vol. 20, №1. – P. 735-746.
30. Maheswaranathan N., Williams A.H., Golub M.D., Ganguli S., Sussillo D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 15696-15705.
31. Li Z., Gurgel H., Dessay N., Hu L., Xu L., Gong P. Semi-Supervised Text Classification Framework: An Overview of Dengue Landscape Factors and Satellite Earth Observation // International Journal of Environmental Research and Public Health. – 2020. – Vol. 17, №12. – P. 4509-4538.
32. Kaewphan S., Hakala K., Miekka N., Salakoski T., Ginter F. Wide-scope biomedical named entity recognition and normalization with

CRFs, fuzzy matching and character level modeling // Database. – 2018. – Vol. 2018 (bay096). – P. 1-10
33. Campos D., Matos S., Oliveira J.L. A document processing pipeline for annotating chemical entities in scientific documents // Journal of Cheminformatics. – 2015. – Vol. 7 (Suppl 1: S7). – P.1-10.
34. Korvigo I., Holmatov M., Zaikovskii A., Skoblov M. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules // Journal of Cheminformatics. – 2018. – № 1. – P. 28.
35. Luo L., Yang Z., Yang P., Zhang Y., Wang L., Lin H., Wang J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition // Bioinformatics. – 2018. – Vol. 34, № 8. – P. 1381-1388.
36. Hemati W., Mehler A. LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools // Journal of Cheminformatics. – 2019. – Vol. 11, № 3. – P. 1-7.
37. Lung P.-Y., He Z., Zhao T., Yu D., Zhang J. Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering // Database.–- 2019. – Vol. 2019 (bay138). – P. 1-8.
38. Capuzzi S.J., Thornton T.E., Liu K., Baker N., Lam W.I., O'Banion C.P., Muratov E.N., Pozefsky D., Tropsha A. Chemotext: A Publicly Available Web Server for Mining Drug–Target–Disease Relationships in PubMed // Journal of Chemical Information and Modeling. – 2018. – Vol. 58, № 2. – P. 212-218.
39. Mao Y., Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank // Journal of Biomedical Semantics. – 2017. – Vol. 8, № 1. – P. 15-24.
40. Пономаренко Е.А., Лисица А.В., Ильгисонис Е.В., Арчаков А.И. Создание семантических сетей белков с использованием PUBMED/MEDLINE // Молекулярная Биология. – 2010. – Т. 44, № 1. – С. 152-161.
41. Vempati U.D., Schürer S.C. Development and Applications of the Bioassay Ontology (BAO) to Describe and Categorize High-Throughput Assays // Assay Guidance Manual / eds. S. Markossian, G.S. Sittampalam, A. Grossman, et al. – Bethesda: Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2004. – P.1045-1069.
42. Hastings J., Chepelev L., Willighagen E., Adams N., Steinbeck Ch., Dumontier M. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web // PLoS ONE. – 2011. – Vol. 6, № 10. – P. e25513.

Материал поступил в редакцию 31.08.20.

Сведения об авторах

БИЗЮКОВА Надежда Юрьевна – лаборант Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича, студентка шестого курса специальности «Медицинская кибернетика» Российского национального исследовательского медицинского университета имени Н.И. Пирогова, Москва
e-mail: nad.smol@gmail.com

ТАРАСОВА Ольга Александровна – кандидат биологических наук, научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: olga.a.tarasova@gmail.com

РУДИК Анастасия Владимировна – кандидат биологических наук, старший научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: rudik_anastassia@mail.ru

ФИЛИМОНОВ Дмитрий Алексеевич – кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: dmitry.filimonov@ibmc.msk.ru

ПОРОЙКОВ Владимир Васильевич – доктор биологических наук, кандидат физико-математических наук, член-корреспондент РАН, профессор, главный научный сотрудник, заведующий отделом биоинформатики Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича
e-mail: vladimir.poroikov@ibmc.msk.ru;
vvp1951@yandex.ru

ВИНИТИ РАН

Центр научно-информационного обслуживания

Информационные услуги, предоставляемые ЦНИО ВИНИТИ РАН:

– проведение тематического поиска и консультации поисковых экспертов;
– подготовка списков научной литературы;
– подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
– библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
– информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

– фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
– базами данных и Интернет-ресурсами: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
– полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНИТИ РАН www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНИТИ http://catalog.viniti.ru** и **служба электронной доставки документов.**

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНИТИ РАН.

Подробную информацию Вы можете получить:
Адрес: 125190, Россия, г. Москва, ул. Усиевича, 20, ВИНИТИ РАН;
Телефоны: 499-155-42-17, 499-155-42-43;
E-mail: cnio@viniti.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

http://www.udcc.ru