Московский физико-технический институт (государственный университет)

Факультет биологической и медицинской физики

Кафедра кафедра молекулярной и трансляционной медицины

Диссертация допущена к защите

зав. кафедрой

_____ Лазарев В.Н.

«_____» _____ 2016 г.

Выпускная квалификационная работа на соискание степени МАГИСТРА

Тема: **Автоматический анализ текстов для выявления отношений между бактериями микробиоты кишечника человека, питанием и заболеваниями**

Направление: 010900 – Прикладные математика и физика

Магистерская программа: 010982 – Физико-химическая биология и биотехнология

Выполнил студент гр. 0112 _____ Ярыгин К. С.

Научный руководитель,

к. б. н. _____ Лазарев В.Н.

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России

Москва – 2016

Оглавление

1. Список сокращений

ЖКТ — Желудочно-кишечный тракт

ИАТ — Интеллектуальный анализ текстов

ИАБТ — Интеллектуальный анализ биомедицинских текстов

КЖК — Короткоцепочечные жирные кислоты

MEMM, ММЭМ — Maximum entropy markov models, Марковские модели максимальной энтропии

SVM, МОВ — Support vector machine, Метод опорных векторов

NLP, ОЕЯ — Natural language processing, Обработка естественных языков

NER, РИС — Named entity recognition, Распознование именнованных сущностей

CRF, УСП — Conditional random fields, Условные случайные поля

CSD — Complementary structures in disjoint literatures, Комплементарные структуры в непересекающихся литературах

GO — Gene ontology, Онтология генов

MeSH — Medical subject headings, Медицинские предметные рубрики

NGS — Next-Generation sequencing, Секвенирование нового поколения

UMLS — Unified medical language system, Унифицированный язык медицинских систем

2. Введение

Микробные сообщества, населяющие различные части кишечника человека, могут оказывать влияние на здоровье своего хозяина. В здоровом организме они способствуют пищеварению, расщепляя не переваренные человеком компоненты пищи и синтезируя различные нутриенты и витамины. Негативные эффекты, оказываемые кишечными микробами, включают в себя различные воспаления и инфекции, участие в возникновении и развитии болезней желудочно-кишечного тракта (ЖКТ), а также диабета и ожирения. В последнее десятилетие был достигнут огромный прогресс в методах определения состава кишечной микробиоты и роли бактерий в кишечном метаболизме. Стало понятно, что питание более всего остального определяет состав микробиоты в краткосрочной и долгосрочной перспективе, что открывает возможности манипулировать состоянием микробиоты меняя диету. Также стала ясна роль межиндивидуальных различий в составе микробиоты, которые вероятно служат причиной различных ответов на прием лекарственных средств или смену диеты. Достижение более точного понимания метаболических ролей различных бактерий в кишечнике, их взаимодействия друг с другом и эффектов, которые это взаимодействие оказывает на человека, очень важно для лечения и предотвращения возникновения болезней с микробиотной компонентой в этиологии.

К настоящему времени накоплено огромное количество научных данных о микробиоте кишечника человека. На данный момент база данных научной литературы Pubmed Central индексирует около 20,000 статей на тему “Микробиота кишечника человека” и с каждым годом это количество увеличивается экспоненциально. Учитывая это обстоятельство, следить за релевантными публикациями становится все сложнее, а поиск нужной информации вручную во всех накопленных источниках и сопоставление ее с результатами экспериментов — процесс, занимающий много времени. Для автоматизации же процесса поиска с помощью компьютера.

недостаточно просто загрузить в него грамматику и словарь. Компьютер, как и человек, должен обладать специализированным знанием чтобы понимать текст. Научная область, занимающаяся обучением компьютера понимать написанный человеком текст называется обработкой естественных языков (ОЕЯ, *англ.*: natural language processing, NLP). Интеллектуальный анализ биомедицинских текстов (ИАБТ, *англ.*: biomedical text mining) — ее подраздел, посвященный обработке научных текстов по

биологии, химии и медицине. Основная цель ИАБТ — вычленить какую-либо биологическую информацию из текста, написанного человеком на естественном языке. В настоящее время существует множество систем ИАБТ, успешно применяемых для нахождения белок-белковых взаимодействий [1], ген-белковых взаимодействий [2], определение мишеней лекарственных препаратов [3] и многое другое.

Целью данной работы является создание алгоритма ИАБТ для сбора информации о связях между бактериями, болезнями и питанием из научной литературы. Для этого были поставлены следующие задачи:

- Разработать алгоритм для нахождения информации о биологических связях между бактериями и болезнями, а также между бактериями и нутриентами в произвольном тексте.
- Применить разработанный алгоритм на текстах научных публикаций на тему “Микробиота кишечника человека”.
- Провести валидацию метода на основе полученных результатов.

3. Литературный обзор

3.1. Микробиота кишечника человека

Большинство многоклеточных организмов живут в тесной связи с окружающими микробами, и человек не является исключением. Человеческий организм населен большим количеством разнообразных бактерий, архей, вирусов и одноклеточных эукариотов. Микроорганизмы, сосуществующие вместе со своим хозяином, называются микробиотой. Состав микробиоты и ее влияние на организм хозяина стали объектом пристального изучения в последний десяток лет, с появлением технологий NGS (*англ.*: next-generation sequencing, секвенирование нового поколения). В данный момент бактерии остаются самым изученным компонентом микробиоты, в то время как роль вирусов, архей и эукариот менее ясна. По оценкам, микробиота человека содержит около 10^{13}–10^{14} бактериальных клеток, примерно столько же, сколько содержится в самом человеческом теле [4]. Микробиотой обладают практически все поверхности человеческого тела, которые так или иначе имеют контакт с внешней средой: микробы покрывают кожу, желудочно-кишечный, мочеполовой и дыхательный тракт. Самой заселенной средой является желудочно-кишечный тракт (ЖКТ): 70% всех микробов, населяющих человеческое тело, содержатся в толстом кишечнике [5]. Это вызвано тем, что кишечник обладает большой площадью поверхности ($\sim 200 м^2$) [6], а также содержит множество веществ, которые могут использоваться микробами в качестве питания.

Большинство бактерий, проживающих в кишечнике взрослого человека — это неспорообразующие анаэробы. Наиболее представлены среди них бактерии видов *Bacteroides*, *Bifidobacterium*, *Eubacterium*, *Clostridium*, *Lactobacillus* и *Fusobacterium*, а также различные грамположительные кокки [7]. Кишечные микроорганизмы обладают способностью метаболизировать доступные субстраты, которые могут иметь как внешнее происхождение (пища), так и внутреннее (муцин). Основные субстраты, доступные кишечным бактериям, это крахмал и пищевые волокна, а также другие углеводы, такие как олигосахариды и не поглощаемые человеком сахара. Что касается белков, кишечные бактерии могут использовать ферменты хозяина, муцин, а также отслоившиеся клетки эпителия.

Основываясь на их метаболической активности и продуктах ферментации, ки-

шечные бактерии могут быть поделены на полезных и потенциально патогенных. Положительные эффекты на здоровье со стороны микробиоты включают в себя помощь в пищеварении и усвоении пищи, иммуностимулирующее действие, синтез витаминов, подавление роста потенциальных патогенов и уменьшение холестерина [8–10]. Примерами полезных бактерий, особенно важных для человека, могут служить бактерии видов *Lactobacilli* или *Bifidobacteria*, которые помогают переваривать лактозу людям с ее непереносимостью [11], увеличивают устойчивость к инфекциям [12], а также помогают справляться с воспалительными процессами в кишечнике [13]. Отрицательными же эффектами могут являться синтез микробиотой канцерогенных веществ и токсино [14], диарея [15], запоры [16] или даже болезни печени [17], а так же кишечные инфекции. Примерами патогенных бактерий являются *Pseudomonas aeruginosa*, вызывающая инфекционную диарею [18], или *Clostridium difficile*, служащая причиной псевдомембранозного колита [19]

Микробиота и питание

Микробный состав кишечной микробиоты взрослого человека стабилен и мало изменяется со временем. Влияние определенных пищевых компонентов было исследовано в нескольких работах [20; 21], но, несмотря на то, что многие исследования показали наличие реакции микробиоты на определенный пребиотик, только в нескольких были отмечены временные изменения всего микробного состава [22–24]. Похоже, что состав микробиоты зависит больше от индивидуальных особенностей, чем от диеты [22]. С другой стороны, диета все же оказывает влияет на относительную представленность различных групп бактерий в микробиоте. При смене диеты (равно как и при приеме антибиотиков) изменения в представленности бактерий происходят довольно быстро (в течение нескольких дней) и с такой же скоростью возвращаются к начальным значениям при возврате к привычной диете [22].

Метаболическая активность бактерий в кишечной микробиоте может иметь сильные последствия для хозяина, которые могут быть как положительными, так и пагубными. Метаболизм анаэробных микробных сообществ являет собой сложную систему, где синтрофия (симбиотическое сосуществование) является широко распространенным явлением [25]. Только в малой части случаев можно ассоциировать определенные продукты метаболизма с одной или несколькими бактериями. Обычно метаболи-

ты производятся и поглощаются множеством различных членов сообщества. Состав микробиоты, равно как и доступные для нее субстраты, определяют финальные продукты метаболизма сообщества.

У человека нет ферментов, необходимых для расщепления многих структурных углеводов, содержащихся в употребляемых в пищу растениях, и поэтому они непереваренными доходят до толстого кишечника, где служат субстратом для бактерий. Основные формы углеводов, доступных для кишечных бактерий — это резистентный крахмал и различные некрахмальные поли- и олигосахариды. Расщепление различных структурных углеводов в кишечнике крайне специализированно, например, основным расщепителем крахмала является *Ruminococcus bromii*, а целлюлозы — *Ruminococcus champanellensis*. Продуктом ферментации углеводов кишечными бактериями являются короткоцепочечные жирные кислоты (КЖК), в основном ацетат, пропионат и бутират. Бутират является основным источником для клеток эпителия кишечника, пропионат транспортируется к печени, где участвует в глюконеогенезе, а ацетат попадает в кровоток и используется в липогенезе [26]. Также в результате ферментации неперевариваемых углеводов некоторыми бактериями выделяются вещества, которые могут оказывать антивоспалительное и антиоксидантное действие [27]. В противовес ферментации углеводов, бактериальный метаболизм пищевых жиров и белков в толстом кишечнике может привести к синтезу вредных метаболитов, которые могут быть вредны для человека, например, нитрозаминов [14]. Также микробиота может использовать эндогенные субстраты, такие как человеческие ферменты, клетки эпителия кишечника и муцин (например, *Akkermansia mucinophila* — главный потребитель муцина в кишечнике человека [28])

Микробиота и болезни

Бактериальный состав микробиоты кишечника может флуктуировать со временем, например, при инфекциях или при приеме антибиотиков. Различные болезни также могут оказывать сильное влияние на представленность бактерий в микробиоте. Сейчас для множества болезней доказана их непосредственная связь с с состоянием микробиоты кишечника. Среди них — заболевания с аутоиммунной компонентой, такие как целиакия [29], сахарный диабет 2-го типа [30] и ожирение [31]; воспалительные заболевания кишечника [32] (такие, как болезнь Крона или язвенный колит) и

хроническая диарея [15]. Несмотря на установленную связь между некоторыми болезнями и микробиотой, зачастую роль кишечных бактерий в возникновении и развитии болезней ясна не до конца.

То, насколько важна сбалансированная микробиота, особенно заметно, когда речь идет о последствиях приема антибиотиков. Некоторые исследования показали негативный эффект антибиотиков на кишечную микробиоту [33–35]. Отрицательные последствия приема антибиотиков длятся долгое время даже после прекращения их употребления, что свидетельствует о вредоносных эффектах, вызванных дисбалансом в микробном сообществе. Одно из наиболее известных последствий такого дисбаланса — антибиотик-ассоциированная диарея, вызванная усиленным размножением патогенного вида *Clostridium difficile* [36]. Нормальная микробиота подавляет рост патогенов, конкурируя с ними за субстраты, вырабатывая бактериоцины или разрушающие токсины протеазы. Антибиотики разрушают этот сложный комплекс взаимодействий между микробами, позволяя патогенам резко увеличивать свою численность. Поэтому лечение антибиотик-ассоциированной диареи, вызванной *Clostridium difficile*, с помощью ванкомицина и пробиотических дрожжей является более эффективным по сравнению с лечением только ванкомицином [37].

Несмотря на множество защитных механизмов, представленных в слизистой кишечника, человек все же может подвергнуться заражению различными кишечными патогенами (как вирусами, так и бактериями). Колонизация слизистой кишечника патогенными бактериями влечет за собой воспалительный ответ со стороны организма хозяина с целью сопротивления заражению. Однако воспаление также производит обратный эффект, уменьшая жизнеспособность кишечной микробиоты и позволяя патогену занять освободившиеся ниши. Множество патогенных бактерий используют такой воспалительный подход для максимизации своего инфекционного потенциала: *Citrobacter rodentium* [38], *Salmonella enterica* [39], *Helicobacter trognotum* [40]. В то же время воспаление также дает возможности для роста оппортунистических патогенов, таких как *Enterobacteriaceae* [38].

3.2. Интеллектуальный анализ текстов

Научная литература предоставляет огромное количество информации для исследователей. Эта информация может быть использована для построения гипотез,

для интерпретации результатов экспериментов или для овладения знаниями в той или иной научной области.

Существует множество баз данных, в которых содержится биомедицинская научная литература. Одна из самых крупных из них — PubMed — в данный момент содержит более 25 миллионов ссылок на научную литературу из базы данных MEDLINE, биомедицинских журналов и книг (http://www.ncbi.nlm.nih.gov/pubmed). Количество ежегодно публикуемых статей растет экспоненциально [41], вместе с ним растет и количество различных экспериментальных данных, получаемых научными лабораториями во всем мире.

Учитывая экспоненциальный рост количества публикаций, следить за релевантными для ученого публикациями становится все сложнее, не говоря о публикациях из смежных областей. Вычленение нужной информации из баз данных научной литературы и сопоставление ее с результатами экспериментов — процесс, занимающий много времени и требующий аккуратного подбора ключевых слов и составления правильных запросов. Таким образом, ученому зачастую не удается проанализировать все релевантные статьи на определенную тему.

Интеллектуальный (или машинный) анализ текстов (ИАТ, *англ.* text mining) — это автоматическая обработка текста и его анализ, основанные на методах машинного обучения и обработки естественного языка. В настоящее время ИАТ применяется учеными и исследователями для анализа и обработки научной литературы. С помощью ИАТ становится возможным решить множество различных задач — начиная от выявления белок-белковых взаимодействий [1; 42] до построения базы данных для какой-либо специфической научной области [43].

Из-за того, что текст на естественном языке является весьма неоднородным источником информации, автоматическое вычленение из него необходимой информации — нетривиальный процесс. К настоящему времени ИАТ развился в сложную специализированную дисциплину, в которой обработка текста и машинное обучение сопряжены с анализом данных из различных биологических, химических и медицинских баз данных.

Основная цель ИАТ — вычленить необходимое знание, находящееся в тексте, и представить его в явном и кратком виде; шире — получить новую, ранее неизвестную информацию в результате анализа множества фактов, разбросанных в релевантной

литературе. К сегодняшнему моменту ИАТ, используя факты из баз данных и медикобиологических текстов, успешно применяется для определения молекулярных причин заболеваний [44], ген-белковых [2] и белок-белковых взаимодействий [1; 42], аннотации экспрессии генов [45], определения мишеней лекарственных препаратов [3] и многого другого.

Общая схема анализа

Можно сказать что основная цель ИАТ — добыть из неструктурированного текста, написанного на естественном языке, не выраженные явно знания и представить их в явной форме. Обычно весь процесс ИАТ биомедицинской литературы можно разбить на четыре части:

1. Информационный поиск (*англ.* Information retrieval) — отбор релевантных задаче текстов для последующего анализа
2. Извлечение информации (*англ.* Information extraction) — системы извлечения информации используются для извлечения из анализируемого текста информацию определенного типа. Особо выделяются два подтипа этой задачи:
   - Распознавание именованных сущностей (*англ.* Named entity recognition) — поиск в тексте слов-имен биологических объектов. Это могут быть названия белков, генов, а так же слов, указывающих на них
   - Распознавание отношений (*англ.* Relation extraction) — поиск в тексте отношений между найденными сущностями.
3. Генерация гипотез (*англ.* Hypothesis generation) — системы генерации гипотез делают какие-либо выводы основываясь на данных, полученных при анализе текста

Различные системы ИАТ в зависимости от решаемой задачи могут включать как все вышеперечисленные пункты, так и сосредотачиваться только на одном из них (например распознавание сущностей).

Информационный поиск

Информационный поиск — первый шаг ИАТ, цель которого выбрать набор документов, который потом будет подвергнут анализу. В большинстве своем выбор обуславливается научной областью. Например, если целью исследования служит анализ

причин определенной болезни, то первым шагом будет получение всех биомедицинских текстов (книги, краткие аннотации или полные тексты статей), касающихся ее. Это может быть сделано с помощью специализированных поисковых систем.

Такие системы как PubMed (http://www.ncbi.nlm.nih.gov/pubmed) или PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/) используют для поиска систему логических сочетаний запросов (например "ген_1 AND болезнь_2"). PubMed извлекает документы из базы данных MEDLINE, но предоставляет только аннотации научных статей, PubMed Central предоставляет доступ к полным текстам статей, но только к тем, к которым издательства предоставили свободный доступ.

Также существуют системы, которые не основываются на системе соответствий запросам. Например GoPubMed [46; 47] (http://www.gopubmed.com/web/gopubmed/) классифицирует аннотации статей, используя термины из GO (Gene Ontology) и MeSH (Medical Subject Headings) и организует документы в иерархическую структуру, в которой каждому документу присвоены один или несколько "концептов". GoWeb [48](http://gopubmed.org/web/goweb/) комбинирует веб-поиск на основе ключевых слов, поиск по онтологиям и ИАТ. Textpresso [49] (http://www.textpresso.org/) использует для поиска свою собственную онтологию биологических конецептов (клетка, ген и т.д.) и их отношений. iHOP [50] (http://www.pdg.cnb.uam.es/UniPub/iHOP/) конвертирует статьи из PubMed в граф из генов и белков, делая поиск статей более интуитивным для человека.

Распознавание именованных сущностей

Документы, прошедшие отбор могут быть проанализированы на предмет вхождения в них специфических слов. Этот шаг называется распознавание именованных сущностей (РИС, *англ.*: named entity recognition, NER). Именованная сущность — ключевое слово или набор слов, обозначающий определенный предмет или биологический концепт (это может быть название гена, белка, болезни и т.д.) В ходе РИС такие ключевые слова, найденные в тексте, связываются с концептами, которые в свою очередь, могут упоминаться в тексте под множеством имен. Например, упоминание определенного лекарства должно быть найдено в тексте не только по тривиальному названию химического вещества, но и по синонимам в различных номенклатурах, его торговым названиям и т.д.; поиск упоминаний бактерий должен включать пол-

ное название вида, его сокращенное и устаревшие названия.

Идентификация именованных сущностей Системы распознавания именованных сущностей можно разделить на три большие группы, в зависимости от подхода, который в них применяется [51].

Первая из них — поиск, основанный на правилах (*англ.*: rule-based approach) [52; 53]. Такие методы используют различные характеристики слов и проверяют, удовлетворяют ли он определенным правилам. Характеристиками слова может служить вхождение в него специальных символов, чисел и букв различного регистра (актуально в случае поиска имен генов и белков), а так же его контекст — части речи и грамматические зависимости. Чаще всего поиск, основанный на правилах, используется в связке с поиском, основанном на словаре.

Вторая группа — поиск, основанный на словаре (*англ.*: dictionary-based approach). Данные системы распознавания именованных сущностей основываются на поиске слов, входящих в заранее составленный словарь терминов. При нахождении сущности она связывается с подходящим термином в какой-либо из баз данных. Обычно словари для поиска составляются на основе различных медикобиологических баз данных (например UMLS, GO, ChemID). Системы, основанные на простом прямом поиске по словарю обладают весьма маленькой чувствительностью из-за высокой вариативности написания имен сущностей [54], поэтому были созданы системы, использующие нечеткий поиск [55; 56] и даже выравнивание имен генов друг на друга [57]. В данный момент подходы, основанные на поиске по словарю, остаются наиболее популярными для аннотации медико-биологической литературы, несмотря на возрастающую популярность машинного обучения в задачах РИС.

Третья группа — системы, использующие машинное обучение (*англ.*: machine learning-based approach). Появление таких методов распознавания именованных сущностей стало возможно благодаря постоянно увеличивающемуся количеству уже размеченных текстов (такие совокупности текстов называются корпусами). Наиболее популярные алгоритмы для РИС в биомедицинской литературе — метод опорных векторов (МОВ, *англ.*: Support Vector Machine, SVM) [58; 59], марковские модели максимальной энтропии (ММЭ, *англ.*: Maximum Entropy Markov Models, MEMM) [60] и условные случайные поля (УСП; *англ.*: Conditional Random Fields, CRF) [61–63]. Несмотря на то, что алгоритмы машинного обучения имеют хорошие значения точно-

сти и чувствительности, их эффективность полностью зависит от качества и состава обучающей выборки.

В настоящее время высокоэффективные системы распознавания именованных сущностей совмещают в себе все перечисленные методы и используют различные алгоритмы для пред- и пост-обработки текстов и результатов.

Разрешение полисемии Одна из серьезных проблем распознавания именованных сущностей — разрешение неоднозначности в том случае, если термин имеет несколько смысловых значений (полисемия). Например, "*Casp1*" может означать как ген, так и соответствующей ему белок, "mice" может означать в зависимости от контекста или "мышь" или название белка "*MicE*". "*Streptococcus pneumoniae*" может быть ложно определено как название болезни, а "*Streptococcus pneumoniae* 19A serotype" может быть принято за "*Streptococcus pneumoniae*". Системы разрешения полисемии обычно обучены на огромном массиве энциклопедических данных (например статьи из Википедии) и анализируют контекст, в котором упомянута сущность [64–66].

Извлечение отношений

Большинство задач интеллектуального анализа текстов требуют большего, чем просто нахождение в тексте различных сущностей, и включают в себя поиск отношений, связывающих эти сущности. В самом простом случае отношения между биомедицинскими сущностями бинарны и включают в себя только пару сущностей и зависимость между ними. Например, в предложении "*glnAP2* может быть активирован *NifA*" глагол "активирован" определяет тип отношения, слова "*glnAP2*" и "*NifA*" являются его аргументами. Однако биологические отношения могут быть сложны и включать в себя три и более сущности и несколько связей разных типов. Например предложение "стероид эффективно ингибирует фосфориляцию тирозина у *STAT6*" описывает два биологических отношения, одно из которых определяет существительное "фосфориляция", а другое, сложное отношение, определяемое глаголом "ингибирует" принимает в качестве аргументов сущность "стероид" и первое отношение. Часто отношения между биологическими сущностями также называют "событиями".

Главная цель извлечения отношений — найти в тексте зависимость между сущ-

ностями и определить ее тип. Искомые отношения могут быть как простыми и определенными (как в случае поиска регуляторных отношений между генами), так и сложнокатегоризируемыми, включающими любой тип биомедицинской связи.

В медицине и биологии извлечение отношений имеет множество областей применения. В настоящее время основной задачей извлечения отношений является нахождение белок-белковых взаимодействий, которые играют ключевую роль в структурной и функциональной организации клетки. Другие области, где задействовано извлечение отношений это нахождение взаимодействий между белками и их сайтами связывания, белками и точечными мутациями, генами и болезнями, лекарствами и их мишенями и многое другое [67].

Методы извлечения отношений могут быть разделены на 3 категории: методы, использующие со-встречаемость терминов; методы, основанные на правилах и методы, использующие машинное обучение [68]. В последние время было разработано множество методов извлечения отношений, которые используют гибридные подходы, сочетающие два или больше метода для достижения большей точности и чувствительности при анализе предложений сложной конструкции, часто встречающихся в биомедицинской литературе.

Методы на основе со-встречаемости Анализ со-встречаемости сущностей является простейшим способом нахождения зависимостей. Если сущности часто встречаются вместе, в одном предложении, абзаце или статье, то велика вероятность что они могут быть так или иначе связаны. Например, со-встречаемость терминов “ретинол-связывающий белок 4 ($RBP4$)” и “инсулинорезистентность” в аннотациях статей предполагает функциональную связь между белком и болезнью. Для оценки достоверности используется относительная частота встречаемости двух терминов. Методы, основанные на со-встречаемости легко реализуемы, но, к сожалению, очевидно не способны предоставить информацию о типе отношения, а потому обладают гораздо более низкой точностью по сравнению с другими методами.

Методы на основе правил Для методов, основанных на правилах, используются предопределенные шаблоны, находящие отношения и определяющие их принадлежность к тому или иному типу. Такие правила задаются вручную или высчитываются автоматически из размеченных корпусов. Например, в работе Пру [69] для поиска ген-

генных взаимодействий были использованы вручную сконструированные паттерны. Пример такого паттерна: “[продукт гена] играет роль [модификатор] [ген]”, описывающий определенный сценарий взаимодействия. Такой паттерн способен найти фразу “белок *Eg1* играет роль репрессора *BicD*”, вычленив событие — репрессия, и два его аргуметна: *Eg1* и *BicD*. В PPInterFinder [70], веб-сервисе для нахождения человеческих белок-белковых взаимодействий в аннотациях научных статей искали паттерны в форме “[белок] * [отношение] * [белок]”, “[белок] * [белок] * [отношение]” и “[отношение] * [белок] * [белок]”. Правила могут использовать синтаксическую структуру предложения, морфологию или части речи входящих в него слов. Для автоматического вычленения информации такого рода используются алгоритмы обработки естественных языков (ОЕЯ, *англ.*: natural language processing, NLP). ОЕЯ основана на априорном знании о структуре языка и том, каким образом специфическая биомедицинская информация упоминается в литературе для трансформации предложений, написанных на естественном языке, в структуру, пригодную для автоматической обработки. Анализ таких структур позволяет добиться лучших результатов, чем простой поиск линейных паттернов. В области нахождения зависимостей огромный прогресс был обеспечен использованием таких структур как грамматика составляющих (*англ.*: constituent grammar) и грамматика зависимостей (*англ.*: dependency grammar) [71].

Грамматика составляющих представляет предложение как иерархическую структуру, дерево, листья которого являются единичными словами из этого предложения. Поддеревья, корнями которых служат узлы графа, включающие в себя более одного слова, называются непосредственными составляющими (*англ.*: immediate constituent). В таком виде сложные грамматические единицы складываются из нескольких более простых и не пересекающихся единиц. Сложные грамматические единицы, включающие в себя группу слов, могут принадлежать одной из шести категорий:

1. именная группа — возглавляется существительным
2. группа прилагательного — возглавляется прилагательным
3. наречная группа — возглавляется наречием
4. предложная группа — возглавляется предлогом
5. глагольная группа — возглавляется глаголом
6. предложение

Ниже на рисунке 1 приведен пример дерева грамматики составляющих.

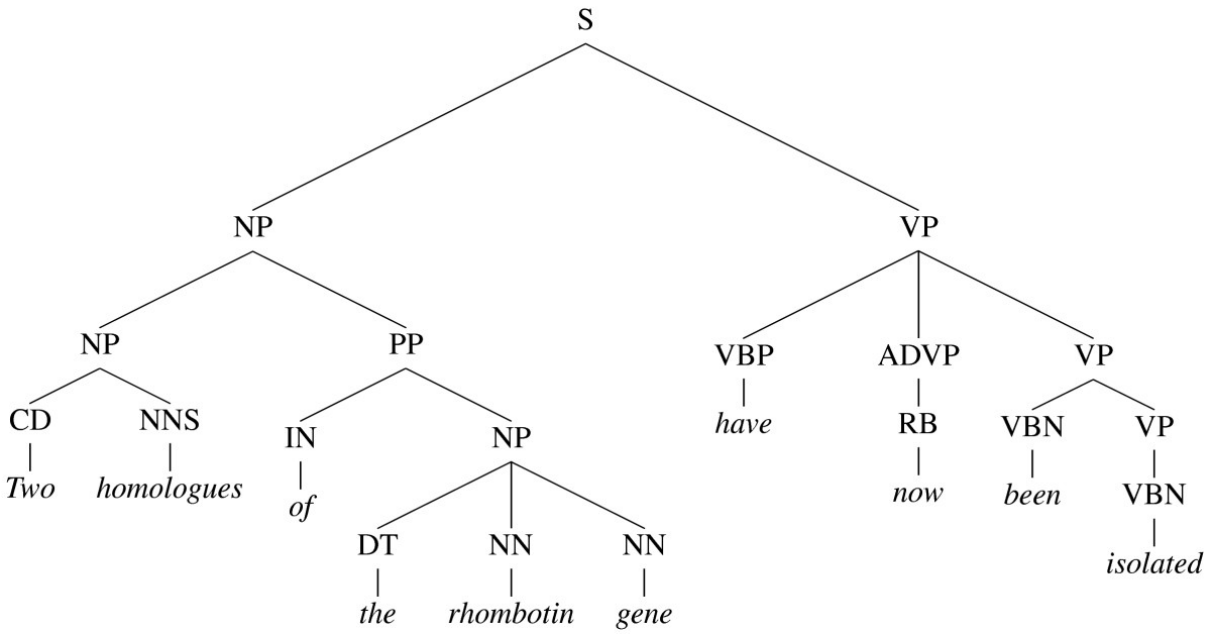

Рис. 1. Дерево грамматики составляющих для предложения “Two homologues of the rhombotin gene have now been isolated” [72].

В отличии от грамматики составляющих (которая более популярна скорее в теоретической лингвистике) грамматика зависимостей гораздо чаще используется в прикладных задачах обработки естественного языка. Грамматика зависимостей представляет предложение в виде графа, узлами которого являются слова из этого предложения, а ребрами — связи между этими вершинами. Связь здесь рассматривается как бинарное асимметричное синтаксическое отношение между двумя словами словами. На рисунке 2 представлен пример такого графа. Ниже на рисунке 2 приведен пример дерева грамматики составляющих.

Множество современных систем извлечения отношений используют грамматику зависимостей. Например RelEx [73] основана на анализе графа зависимостей с помощью набора простых паттернов, использующих именные группы, части речи слов и поиск отрицаний. RelEx фокусируется на отношениях между генами и белками и использует набор из нескольких паттернов взаимодействия для нахождения таковых в тексте. TEES [74; 75] — система извлечения отношений, фокусирующаяся на событиях между белками и генами. TEES использует характеристики графа зависимостей для обнаружения события и метод опорных векторов для его классификации. BExtract [76] основана на анализе кратчайших путей в графе зависимостей между

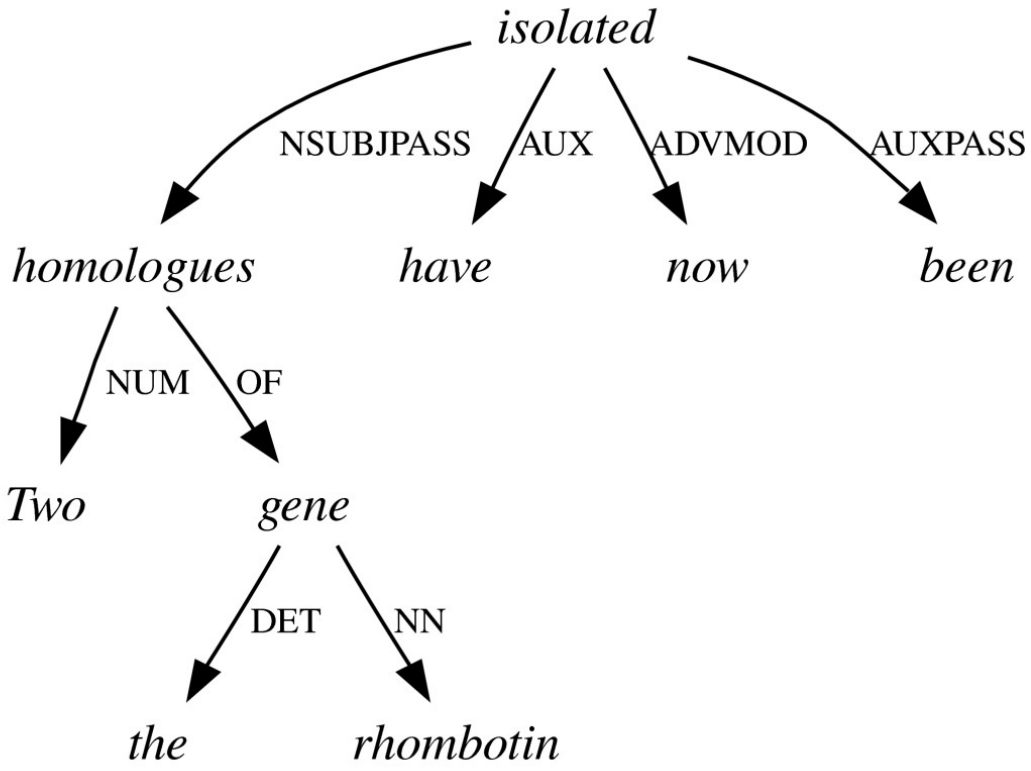

Рис. 2. Граф грамматики зависимостей для предложения "Two homologues of the rhombotin gene have now been isolated" [72].

сущностями и ключевыми выражениями.

Методы на основе машинного обучения Методы, использующие машинное обучение, воспринимают задачу извлечения отношений как задачу классификации. В таких методах какой-либо классификатор (обычно метод опорных векторов) использует обучающую выборку и основываясь на ней определяет тип отношений. Обычно перед использованием машинного обучения предложения подвергаются анализу: определению частей речи и синтаксическому разбору. Далее получившиеся характеристики используются в качестве признаков объектов (предложений) для обучения классификатора. Также входные данные, подаваемые в классификатор, могут быть представлены в сложной форме, такой как граф зависимостей. Основываясь на формате входных данных, методы, использующие обучение с учителем, могут быть разделены на две группы: методы, основанные на признаках (*англ.*: feature-based methods) и ядерные методы (*англ.*: kernel methods) [68] [77] Для методов, основан-

ных на признаках, синтаксические и семантические свойства предложения служат подсказками для решения, содержит ли оно какое-либо отношение и, если содержит, то какого оно типа. Часто используемыми свойствами являются входящие в предложения сущности, части речи слов, кратчайший путь между двумя сущностями в графе зависимостей. В [78] для нахождения белок-белковых взаимодействий авторы взяли следующее множество признаков:

- Признаки отдельных слов: слова, входящие в предложение, их леммы и частоты встречаемости, относительное расположение пары белков внутри предложения
- Кратчайший путь в графе зависимостей: слова-вершины графа и ребра-отношения входящие в кратчайший путь
- Граф зависимостей: матрица связности графа.

Каждая группа признаков была агрегирована в вектор и нормализована. В зависимости от набора признаков, данные методы могут быть разделены на поверхностные и глубокие. Первые анализируют информацию не о всем предложении, а только его части, жертвуя полнотой и глубиной анализа, но являются вычислительно-эффективными. Вторые анализируют структуру всего предложения, чаще достигая лучших значений показателей точности и чувствительности, но требуют больших вычислительных ресурсов. Одним из примеров использования методов, основанных на признаках, может служить система PKDE4J [79]. PKDE4J использует автоматически созданные на основе размеченного корпуса текстов паттерны для анализа графов зависимости. EventMine [80] [81] — сервис ИАТ, использующий машинное обучения для нахождения биологических событий в сочетании с поиском со-представленностей для увеличения точности.

Ядерные методы [82] анализирует предложения как строки, используя ядро для извлечения отношений. Для двух строк x и y строковое ядро (функция, использующая в качестве аргумента строки) вычисляет их схожесть на основании количества их общих подстрок — чем больше общих подстрок, тем больше схожесть между x и y. Каждая строка может быть представлена как вектор в многомерном пространстве, где каждый элемент вектора обозначает либо присутствие (взвешенное, $(0:1]$), либо отсутствие (0) определенной подпоследовательности. В более общем общем представлении x и y могут быть любыми объектами, а функция $K(x, y)$ вычислять их сходство. Так, существуют ядерные функции на подпоследовательностях [83], на деревьях

зависимостей [84], на графах зависимостей [85], на кратчайших путях в графах [86] и на комбинации нескольких объектах из вышеперечисленных [87]. Также есть случаи применения гибридных методов. Например Befree [88] — система для идентификации отношений между медицинскими препаратами, генами и ассоциированными с с ними болезнями — комбинирует ядерные методы с синтаксическим разбором.

В обоих вышеупомянутых группах методах для обучения классификатора нужна обучающая выборка. В биомедицинском ИАТ одним из самых обширных и наиболее тщательно размеченных корпусов является корпус GENIA [89] (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA), состоящий из 2000 аннотаций научных публикаций, полученных по запросам "human", "blood cells" и "transcription factors". Разметка включает в себя части речи, синтаксис, кореференцию (указание разных слов на одну и ту же сущность), биологические концепты и события, клеточную локализация, ассоциации генов с болезнями и метаболические пути. Другие корпуса размеченных аннотаций научных текстов — BioCreAtIve [90–94] и PennBioIE [95]. BioCreAtIve (сокр от англ. "Critical Assessment of Information Extraction systems in Biology"), (http://www.biocreative.org/) — контесты по извлечению отношений в биомедицинских текстах, нацеленные в основном на идентификацию белок-белковых взаимодействий. BioCreAtIve предоставляет размеченный корпуса аннотаций научных публикаций к каждому контесту. Корпус PennBioIE содержит 1100 аннотаций статей, посвященным цитохромам *P-450* и 1157 аннотаций статей, связанных с онкологией. Его разметка содержит разделение на параграфы, предложения, токены, части речи слов, синтаксис и биологические сущности.

Также стоит упомянуть о возможных препятствиях, которые могут возникнуть при анализе текстов на предмет обнаружения отношений и событий. Например, найденное в тексте отношение может не только отражать научный факт, но и быть гипотезой или подвергаться сомнению [96]. Для отделения фактов от гипотез часто требуется ручное курирование результатов анализа. Для того чтобы обойти это проблему некоторые исследования прибегали к статистическому анализу отношений, которые были извлечены из различных литературных источников и баз данных, с последующим их сравнением между собой [3; 97; 98].

Генерация гипотез

В то время, как извлечение отношений фокусируется на поиске отношений, которые явно указаны в тексте, генерация гипотез предпринимает попытки выявить отношения, которые не представлены в тексте, но подразумеваются наличием других, явно указанных отношений.

Практически все работы, посвященные генерации гипотез используют идею, предложенную Д.Р. Свансоном — "комплементарные структуры в непересекающихся литературах" (*англ.*: complementary structures in disjoint literatures, CSD) [99]. Свансон предположил, что огромные базы научной литературы могут позволить сделать открытия с помощью логического связывания различных концептов. Он предложил использовать простой "принцип ABC" — если A оказывает влияние на B, а B оказывает влияние на C, то можно заключить что A может оказывать влияние на C (даже есть A и C ранее не встречались в одном тексте). В своих работах Свансон показал примеры обнаружения новых гипотез, вручную связывая концепты между различными научными статьями. Так, он обнаружил связь, предполагающую что потребление рыбьего жира могло помогать пациентам с болезнью Рейно. Спустя два года клинические испытания установили истинность этой гипотезы [100]. В другой своей работе он отследил 11 опосредованных связей между мигренью и дефицитом магния [99], что также было доказано экспериментально [101; 102].

Свансон применял предложенный им принцип вручную, однако сейчас исследователи автоматизируют этот процесс. Несмотря на то, что принцип ABC прост, его воплощение не является тривиальной задачей. Так как автоматические системы могут генерировать множество гипотез, становится необходим метод оценки их достоверности и оценки доли ложноположительных результатов. Один из способов оценки — возможность воссоздания системой генерации гипотез уже валидированных вручную или экспериментально доказанных фактов (в том числе и открытий Свансона). В работе по поиску новых терапевтических применений талидомида [103] исследователи для валидации гипотез использовали нахождение уже известных фактов, совстречаемость терминов в заголовках и аннотациях статей, а так же ручное курирование результата. В результате другого исследования был обнаружен терапевтические эффекты *Curcuma longa* (куркума) при болезнях сетчатки, болезни Крона и травмах спинного мозга [44; 104]. CoPub Discovery — система, позволяющая поль-

зователю искать скрытые отношения, задавая целевой концепт (A) и список возможных промежуточных концептов (B). Система возвращает все возможные отношения целевого концепта с другими концептами () [105].

Расширением принципа ABC служит профилирование ключевых слов концепта. Например, для определенного концепта (например гена) из научной литературы набирается список ключевых слов, с которыми связан данный концепт. При кластеризации концептов в пространстве ключевых слов, могут быть найдены связанные концепты. Так, профилирование ключевых слов концепта было использовано в онлайн-сервисе Anni [106] и помогло правильно предсказать типы клеток и сигнальные метаболические пути используя данные с микрочипов.

Сейчас доступно множество систем генерации гипотез, которые могут предсказывать мишени лекарств для различных болезней [107], белок-белковые взаимодействия [108; 109] (основываясь на предположении что два белка будут взаимодействовать, если их упоминания в тексте имеют похожий контекст), регуляцию генов [2], фосфорилирование белков [110; 111], отношения между генами [112] и болезнями и ген-генные взаимодействия [113].

3.3. Заключение

Современные методы ИАБТ позволяют анализировать и обрабатывать большие объемы биомедицинских научных текстов, извлекая из них конкретную биологическую информацию в явном и кратком виде, и а также автоматически формулировать научные гипотезы на основе полученной информации. Такие методы становятся особенно востребованными в последнее время из-за экспоненциального роста публикуемых исследований и экспериментальных данных, поскольку ученому становится затруднительно следить за всеми публикациями в своей области. Возможно, что множество важных открытий остаются незамеченными из-за огромного объема накопленной информации. В связи с этим возникает необходимость использования ИАБТ как обязательного инструмента для любого ученого. Однако, применение автоматической обработки текстов остается слабо распространенным явлением в научно-исследовательской деятельности.

К данному моменту разработано огромное количество алгоритмов ИАБТ для различных целей: от нахождения белок-белковых взаимодействий до создания об-

ширных баз данных для какой-либо специфической научной области. Однако, не создано ни одной системы, систематизирующей информацию о влиянии питания на микробиоту кишечника человека и ее роль в развитии различных заболеваний. Таким образом, создание алгоритма, выполняющего такую систематизацию, является важным шагом для дальнейшего изучения взаимного влияния друг на друга микробиоты кишечника человека и ее хозяина и может полезным инструментом для построения гипотез и интерпретации результатов микробиотных исследований.

4. Материалы и методы

4.1. Общая схема анализа

В данной работе исследовались отношения “бактерия-болезнь” и “бактерия-нутриент” в научных публикациях по теме “микробиота кишечника человека”. Тексты статей делились на предложения, каждое из которых анализировалось на предмет содержания в нем названий сущностей из следующих трех групп:

1. Бактерии
2. Болезни
3. Нутриенты

Если предложение их не содержало, либо содержало только названия из одной группы сущностей, оно отбрасывалось и не подвергалось последующему анализу. Отобранные таким образом предложения проходили синтаксический разбор и предобработку: каждое предложение разбивалось на отдельные слова, для слов определялись их части речи, по предложению строился граф зависимостей, в котором находились кратчайшие пути для каждой пары сущностей из разных групп. Используя полученную информацию о предложении оно классифицировалось как не содержащее какого-либо отношения для всех пар сущностей, либо как содержащие (в таком случае также определялся тип этого отношения). Схема анализа представлена на рисунке 3.

4.2. Информационный поиск

Поисковая система Pubmed Central индексирует (по состоянию на 17.12.2015) около 20000 статей по запросу “human gut microbiota”. Из них в открытом доступе находятся 9219 статей. Для анализа использовалось три группы документов:

1. Статьи, которые имеются в открытом доступе в хранилище PubMed Central. Статьи были выкачаны в формате nxml с серверов NBCI через ftp. Количество документов: 9219
2. Статьи, к которым не было отрытого доступа, но копии которых нашлись в онлайн-библиотеке LibGen. Количество документов: 6909
3. Аннотации тех статей, для которых не нашлось полных текстов ни в PubMed Central, ни в LibGen. Количество документов: 6708

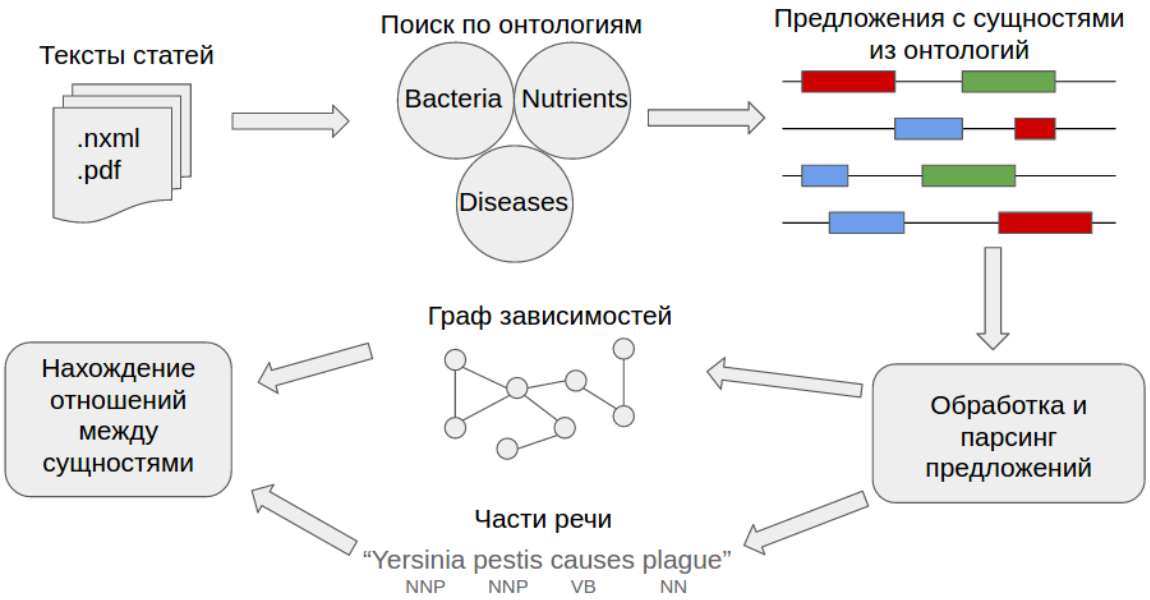

Рис. 3. Схема нахождения зависимостей между бактериями, нутриентами и болезнями в медикобиологичеких текстах.

4.3. Распознавание именованных сущностей

Для поиска именованных сущностей нами было решено использовать поиск, основанный на словаре, как самый распространенный и легко реализуемый. Для каждой из трех групп интересующих нас сущностей (бактерии, болезни и нутриенты) была выбрана собственная онтология или база данных и по ней был составлен словарь терминов.

- **Бактерии.** В качестве основы для составления словаря была выбрана база данных NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy), как наиболее полная и хорошо курируемая. Также NCBI Taxonomy содержит не только номенклатурные названия бактерий на всех таксономических уровнях, но их тривиальные, а так же устаревшие названия. Для составления словаря были взяты только бактерии, населяющие кишечник человека. Список видов кишечных бактерий был взят из работы Д. Ритари [114]. Для каждого вида в словарь были включены его название, названия всех соответствующих ему более высоких таксономических уровней (вплоть до типа), а также их синонимы, устаревшие названия и сокращенные формы в единственном и множественных числах.
- **Болезни.** В качестве основы для составления словаря была выбрана база дан-

ных Disease ontology [115] (http://disease-ontology.org/). Данная онтология содержит названия болезней человека, характеристик фенотипа, связанные с ними медицинский с ними списки медицинских терминов, а также связи между всем вышеперечисленным. Для каждой болезни Disease ontology содержит список синонимов, под которыми она может упоминаться. Для составления словаря были взяты названия болезней, их синонимы, а также всевозможные формы написания в разных регистрах.

- **Нутриенты.** В качестве основы для словаря был взят вручную составленный список веществ, содержащихся в пище. В словарь вошли все вещества из подготовленного списка, их синонимы, номенклатурные и тривиальные названия, а также всевозможные формы написания в разных регистрах.

Для каждого предложения был произведен поиск наличия в нем всех сущностей из всех трех групп. Для каждой найденной сущности сохранялось название, под которым она фигурировала в предложении и ее уникальный идентификатор. Предложение выбрасывалось и не проходило дальнейший анализ, если в нем либо встречались сущности только из одной группы, либо не встречались вовсе.

4.4. Извлечение отношений

Перед извлечением отношений между различными сущностями каждое предложение подвергалось пред-обработке и синтаксическому разбору. В частности, для каждого предложения строился соответствующей ему граф зависимостей и для каждого слова в предложении определялась его часть речи. Из графа зависимостей удалялась информация о направлении связей, граф таким образом становился неориентированным. Если сущность, упоминаемая в предложении, состояла из двух слов или больше, то соответствующие им вершины в графе зависимостей сливались в одну. Также из графа удалялись следующие ребра:

1. Ребра, отражающие связь между придаточной и главной частями в сложноподчиненном предложении
2. Ребра, отражающие сочинение между двумя простыми предложениями в составе сложносочиненного
3. Ребра, отражающие союзную связь между однородными членами предложения, в том случае, если эти члены являются глаголами

В предобработанном таким способом графе для каждой пары сущностей из различных групп находится кратчайший путь между ними. Далее полученная информация (обработанный граф, части речи и кратчайшие пути) использовались для нахождения отношений. Для извлечения отношений между различными сущностями было решено использовать подход, основанный на правилах (паттернах). Данное решение было обусловлено тем, что в настоящее время не существует корпусов с размеченными отношениями между бактериями, нутриентами и болезнями, что исключает применение методов, основанных на машинном обучении. Для определения паттернов было вручную отобрано около двухсот предложений, содержащих отношения между а) бактериями и болезнями, б) бактериями и нутриентами. На основе анализа структуры и свойств предложения и его характеристик (части речи, граф зависимостей и кратчайшие пути) были выбраны несколько характерных паттернов для определения наличия какого-либо отношения в предложении и типа этого отношения. Паттерн представляет собой набор правил, описывающих структуру графа и кратчайшего пути между сущностями.

Паттерны для извлечения отношений "бактерия-болезнь"

Все отношения, которыми могут быть связаны бактерия и болезнь были поделены на четыре группы:

1. Бактерия вызывает болезнь или усугубляет ее течение
2. Бактерия оказывает протективное действие при болезни
3. Болезнь ассоциирована с повышенной представленностью бактерии
4. Болезнь ассоциирована с пониженной представленностью бактерии

- **Паттерн №1**

$$[B] \rightarrow ... \rightarrow [ABUND\_WORD] \rightarrow ... \rightarrow [D]$$
$$\searrow [POS\_QUAL|NEG\_QUAL]$$

Кратчайший путь между сущностями в графе зависимостей должен содержать слово-триггер, служащее для описания представленности ($ABUND\_WORD$ на схеме). Этим словом может являться любое слово, однокоренное с одним из следующих: *presence, abundance, prevalence, amount, occurrence, richness, proportion, incidence, quantity, fraction, portion, percentage, carriage, number*. Вершина, соответствующая

этому слову должна быть связана с вершиной, соответствующей слову, описывающему изменение представленности в большую или меньшую сторону (POS_QUAL и NEG_QUAL на схеме соответственно). Словом, описывающим положительное изменение, может быть любое слово, однокоренное с одним из следующих: *increase, great, enhance, elevated, high, larg*; описывающим отрицательное изменение, соответственно: *reduce, low, diminish, decrease, less, small*. Паттерн служит для описания отношений из 3 и 4 группы.

Пример:

Finally, a ***diminished prevalence*** *and abundance of* ***F. prausnitzii*** *are revealed in the fecal samples of patients with* ***IBD****.*

- **Паттерн №2**

$[B] \rightarrow (0 - 2nodes) \rightarrow [INCREASE|DECREASE] \rightarrow ... \rightarrow [D]$

Кратчайший путь между сущностями должен содержать слово-триггер — слово, однокоренное со словом *increase* либо *decrease*. Расстояние в графе между вершиной, соответствующей слову-триггеру и вершиной, соответствующей названию бактерии, должно быть меньше либо равным трем. Паттерн служит для описания отношений из 3 и 4 группы.

Пример:

Faecalibacterium prausnitzii *is also consistently* ***decreased*** *in human* ***IBD*** *patients and considered an important bacterial group for maintaining microbial homeostasis.*

- **Паттерн №3**

$[B] \rightarrow ... \rightarrow [CAUSE\_WORD|PREVENT\_WORD] \rightarrow (0 - 1nodes) \rightarrow [D]$

Кратчайший путь между сущностями должен содержать слово-триггер, описывающее эффект, оказываемый на болезнь ($CAUSE\_WORD$ и $PREVENT\_WORD$ на схеме). Эффект может быть положительным (протективное действие), в этом случае слово-триггер ($PREVENT\_WORD$) — слово, однокоренное с одним из следующих: *Prevent, attenuate, inhibit, ameliorate, alleviate, counteract, mitigate, protect, suppress, treat*. В случае отрицательного эффекта слово-триггер ($CAUSE\_WORD$) должно быть однокоренным с одним из следующих: *Pcause, provoke, exacerbate, induce, initiate, promote, stimulate, dampen*. Расстояние в графе между вершиной слова-триггера и

вершиной название болезни должно быть меньше либо равно двум. Данный паттерн служит для описания отношений из групп 1 и 2.

Пример:

PThe assessment of gastritis based on histopathological criteria was completely changed after recognition of ***H. pylori*** *as the most common* ***cause*** *of* ***chronic gastritis****.*

- **Паттерн №4**

$$[B] \rightarrow ... \rightarrow [CORELL] \rightarrow ... \rightarrow [D]$$
$$\searrow ?[NEG|INVERS]$$

Кратчайший путь между сущностями должен содержать слово, однокоренное со словом *correlation*. Также оно может быть связано с вершиной, соответствующей слову, однокоренному со словом *inversely* или *negatively*, что будет сигнализировать об отрицательной зависимости. Паттерн служит для поиска отношений группы 3 или 4. В том случае, если выполняется второе условие (связь с *inversely* или *negatively*), то отношение относится к четвертой категории, иначе — к третьей.

Пример:

We found that the ***colitis*** *scores* ***correlated*** *with the abundance of* ***Burkholderia*** *(p=0.0045) and the richness of* ***Lactobacillus*** *group (p=0.006).*

- **Паттерн №5**

$$[B] \rightarrow [D]$$
$$\searrow [CAUSE\_WORD|PREVENT\_WORD]$$

Длина кратчайший пути должна быть 1 (бактерия и болезнь связаны в графе непосредственно). Также вершина, соответствующая названию болезни, должна быть связана с вершиной, соответствующей слову-триггеру ($CAUSE\_WORD$ и $PREVENT\_WORD$ на схеме). Слова-триггеры те же, что и используются в паттерне №3. Данный паттерн служит для описания отношений из групп 1 и 2.

Пример:

Telomere shortening is closely associated with severity of ***H. pylori induced gastritis*** *and CDH1 methylation status.*

Паттерны для извлечения отношений “бактерия-нутриент”

Все отношения, которыми могут быть связаны бактерия и нутриент, были также разделены нами на четыре группы. Для каждой группы был составлен список слов-триггеров, которые могут описывать данное взаимодействие (см Табл. 1).

Таблица 1. Группы взаимодействий и слова-триггеры для каждой из них.

Номер группы	Группа 1	Группа 2	Группа 3	Группа 4
Тип взаимодействия	Бактерия может поглощать нутриент	Бактерия использует в своем метаболизме использует нутриент как субстрат	Бактерия производит нутриент в результате своего метаболизма	Бактерия секретирует вещество
Слова-триггеры, описывающие взаимодействие	*consume, acquire, absorb, ingest, retain, adhere, adopt, assimilate, obtain, eat*	*degrade, utilize, utilise, metabolize, metabolise, ferment, hydrolyze, cleave, catabolize, digest, bind, hydrolyse, lyse, convert, destroy, eliminate, dephosphorylate, exploit, transform, eradicate, dissociate, deactivate, neutralize, dimerize, inactivate, phosphorylate, heterodimerizes, decarboxylate*	*produce, generate, synthesize, synthesise, biosynthesize, biosynthesise, assemble, create, make*	*secrete, excrete, liberate*

Здесь, в отличии от паттернов для нахождения отношений между бактериями и болезнями, любой из паттернов может подойти для описания любой из групп взаимодействий. Тип найденного отношения определяется исключительно группой слова-триггера, найденного в кратчайшем пути.

- **Паттерн №1**

$$[B] \rightarrow ... \rightarrow [VERB\_TRIGGER] \rightarrow (0-1 nodes) \rightarrow [N]$$
$$obj$$

Кратчайший путь между сущностями в графе зависимостей должен содержать вершину, соответствующую слову, однокоренному с одним из слов триггеров, причем это слово должно являться глаголом (*VERB_ TRIGGER* на схеме). Расстояние между этой вершиной и вершиной-нутриентом должно быть равно 1 или 2. Связь, ведущая от слова-триггера к названию нутриента, должна отражать прямое или косвенное дополнение (*англ.*: object)

Пример:

*The **Lactobacillus** strains were investigated for their capability to **assimilate cholesterol** under simulated intestinal conditions.*

- **Паттерн №2**

$$[B] \rightarrow [VBN\_TRIGGER] \rightarrow ... \rightarrow [N]$$
$$\searrow [by]$$

Кратчайший путь должен между сущностями должен содержать вершину, соответствующую слову, однокоренному с одним из слов триггеров, и это слово должно быть причастием прошедшего времени (*VBN_ TRIGGER* на схеме). Вершина-триггер должна быть связана с вершиной, соответствующей слову *by*, а также расстояние между ней и вершиной-названием бактерии должно быть равным 1.

Пример:

***Butyric acid** is one of the metabolites **produced by F. prausnitzii**, which is an important energy source for the intestinal epithelial cells and has an inverse correlation with disease activity.*

- **Паттерн №3**

$$[B] \rightarrow [*] \rightarrow [VBN\_TRIGGER] \rightarrow ... \rightarrow [N]$$
$$\searrow [by]$$

Этот паттерн аналогичен паттерну №2, с той лишь разницей, что между вершиной триггерного слова и вершиной названия бактерии может находиться еще одна вершина (любая), с которой и должно быть связано слово *by*.

Пример:

The amount of ***cholesterol assimilated by*** *the probiotic* ***Lactobacillus*** *strains in terms of a dose of 10 cells was calculated.*

- **Паттерн №4**

$[B] \rightarrow ... \rightarrow [NOUN\_TRIGGER] \rightarrow [N]$

Кратчайший путь должен между сущностями должен содержать вершину, соответствующую слову, однокоренному с одним из слов триггеров, и это слово должно быть существительным ($NOUN\_TRIGGER$ на схеме). Вершина-триггер должна напрямую быть связана с вершиной, соответствующей названию нутриента.

Пример:

The main ***producers*** *of* ***folate*** *are* ***Bifidobacterium bifidum*** *and* ***B. longum*** *subsp.*

После того, как предложение подошло под один из паттернов, в графе зависимостей производится поиск связанных с вершиной, соответствующей слову-триггеру, ребер с тэгом *neg* (от *англ.* negation, отрицание). Если такие ребра есть, то предложение отбрасывается, как содержащее отрицание. Пример такого предложения:

Unlike other enteropathogenic bacteria, ***C. jejuni cannot utilize glucose*** *as a growth substrate since it lacks the glycolytic enzyme phosphofructokinase of the Embden-Meyerhof-Parnas (EMP) pathway.*

Дополнительно, предложение проходит пост-обработку в том случае, если слово-триггер принадлежит второй или третьей группе (вещество участвует в метаболизме бактерии в качестве субстрата или продукта соответственно). Тогда в графе зависимостей ищутся смежные с вершиной-нутриентом узлы, соответствующие предлогам *to*, *into* и *from*, которые могут изменить смысл отношения. Например, для предложения

Another species of great value for industry is ***C. thermocellum strain****, which is an anaerobic thermophile capable of* ***converting*** *waste cellulose* ***into ethanol****.*

парой батерия-нутриент является (*C. thermocellum*; *ethanol*). Предложение подходит под паттерн №1 со словом-триггером *convert* и поэтому отношение сперва отнесено к группе 2 (использование в качестве субстрата). Однако нахождение предлога *into* позволяет отнести отношение к группе 3 (продукт метаболизма).

4.5. Используемые программные пакеты

Весь вычислительный конвейер для проведения анализа был написан на языке python3.4 и доступен в открытом доступе: https://github.com/KonstantinYarygin/relation_extraction/tree/kost. Для лингвистического анализа использовался моду nltk [116]. Синтаксический разбор предложений был сделан с помощью Stanford dependency parser [117].

5. Результаты и обсуждение

Всего во всех текстах трех исследуемых корпусов текстов было найдено 46916 предложений, содержащих пары сущностей “бактерия-нутриент” или “бактерия-болезнь” (Таблица 2).

Таблица 2. Количество предложений с найденными парами сущностей из двух разных групп..

	Полные тексты PubMed Central	**Полные тексты Libgen**	**Тексты аннотаций PubMed Central**	**Все корпуса**
“Бактерия-болезнь”	10634	17504	1778	29916
“Бактерия-нутриент”	7185	9007	808	17000
Все	17819	26511	2586	**46916**

В 46916 предолжениях с обнаруженной со-встречаемостью сущностей было найдено 13084 отношений, зафиксированных паттернами: 8818 отношений между болезнью и бактерией и 4266 отношений между бактерией и нутриентом.

5.1. Отношения между бактериями и болезнями

Среди исследованных текстов было найдено 8818 отношений между болезнями и бактериями, из них 2366 уникальных. Большинство отношений было найдено с помощью паттерна №3 (Рис. 4):

Чаще всего в статьях упоминают бактерии *Escherichia coli*, *Helicobacter pylori* и *Peptoclostridium difficile*, и такие болезни как колит, диарея, синдром раздраженного кишечника (СРК) и ожирение (Рис. 5, Рис. 6)

Самые частоупоминаемые отношения между бактерией и болезнью — *Peptoclostridium difficile* и *Escherichia coli* вызывает диарею, а представленность бактерий типа *Firmicutes* положительно коррелирует с ожирением. На рисунке 7 показаны отношения, которые были найдены максимальное количество раз.

Видно, что некоторые найденные отношения противоречат другим, например в некоторых отношениях *Peptoclostridium difficile* вызывает диарею, в других — вызы-

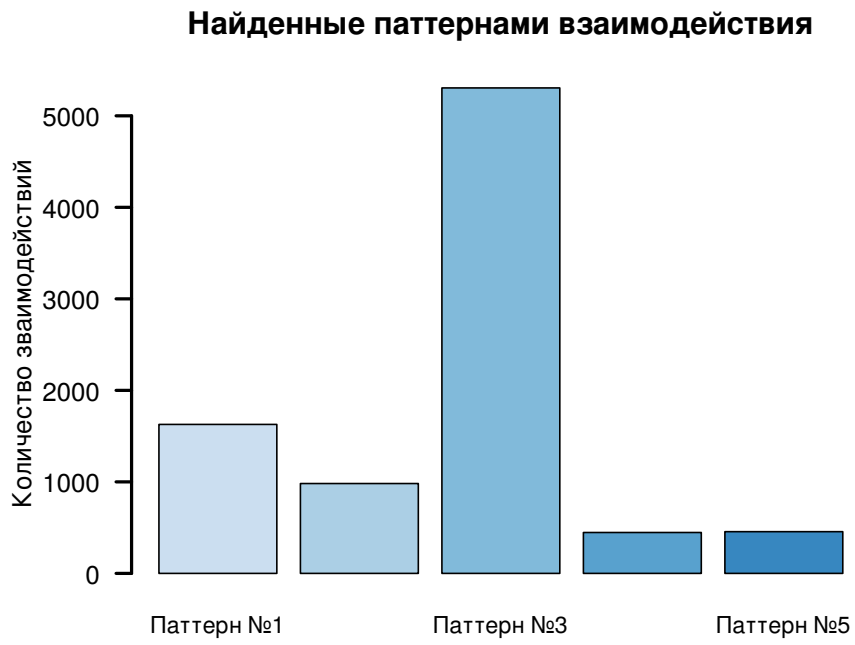

Рис. 4. Количество взаимодействий типа “бактерия-болезнь”, обнаруженных с помощью каждого паттерна.

вает протективное действие. Нахождение вторых — результат ложноположительного срабатывания того или иного паттерна. Например, в предложении

Meta-analysis of probiotics for the prevention of antibiotic associated ***diarrhea*** *and the* ***treatment*** *of* ***Clostridium difficile*** *disease.*

паттерн №3 идентифицировал протективное действие *Clostridium difficile* при диареи.

5.2. Отношения между бактериями и нутриентами

Во всех исследованных текстах было найдено 4266 отношений между бактериями и нутриентами и 2018 уникальных отношений. Самыми результативными оказалить паттерны №4 и №1 (Рис. 8)

Бактерии, которые чаще всего упоминаются в контексте взаимодействий с нутриентами — это *Escherichia coli*, *Bifidobacterium* и *Lactobacillus* (Рис. 9). Стоит заметить, что последние две являются пробиотиками, т.е. бактериями, для которых подтверждено положительное влияние на микробиоту и которые используются в терапевтических целях. Самые упоминаемые нутриенты же это углеводы, в частности целлюлоза, глюкоза и крахмал — основные источники пищи для кишечных бактерий

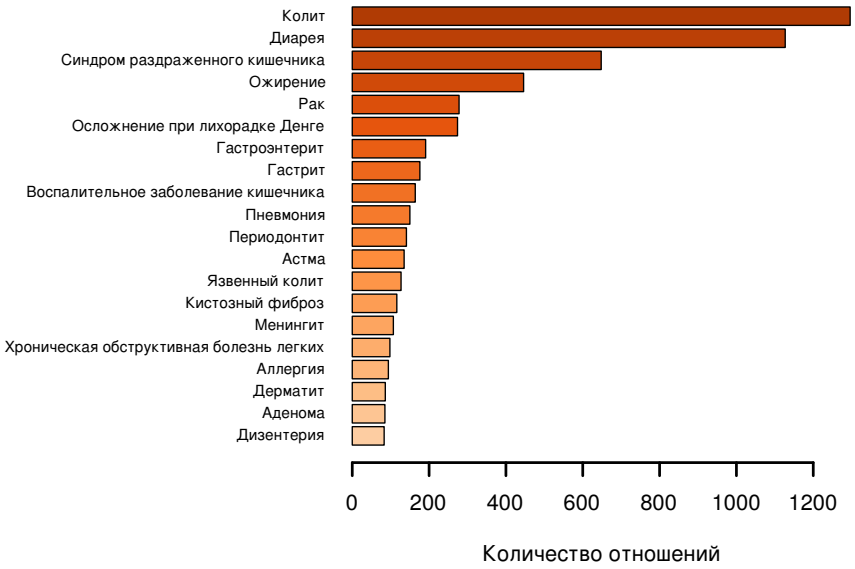

Рис. 5. Болезни, наиболее часто входящие в отношения типа “бактерия-болезнь”.

(Рис. 10).

Самые часто упоминаемые отношения состоят в основном из фактов об использовании углеводов различными бактериями в качестве субстратов (Рис. 11). Однако стоит отметить способность *Pseudomonas aeruginosa* производить альгиновую кислоту и метаболизирование серина и пролина бактерией *Campylobacter jejuni*.

5.3. Валидация результатов

Отношения, полученные в результате работы программы были вручную проверены на предмет содержания найденного факта. Из 229 предложениях, которые были классифицированы как содержащие отношения между бактерией и нутриентом, 193 действительно содержали такие отношения и 36 предложений были классифицированы ошибочно. Так же, из 72 предложений, содержащих отношения между бактерией и болезнью 60 были классифицированы правильно и 12 ошибочно. Итого, точность нахождения связи “бактерия-болезнь” — 0,83, точность нахождения связи “бактерия-нутриент” — 0,84.

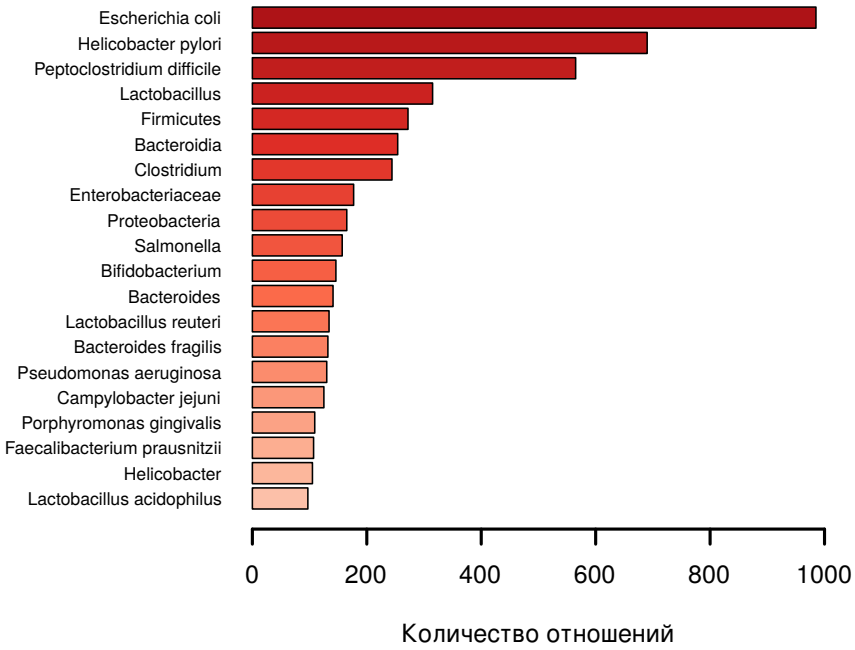

Рис. 6. Бактерии, наиболее часто входящие в отношения типа “бактерия-болезнь”.

6. Выводы

- Разработан алгоритм для нахождения информации о биологических связях между бактериями и болезнями, а также между бактериями и нутриентами в произвольном тексте
- Разработанный алгоритм применен к текстам научных публикаций на тему “Микробиота кишечника человека”. В частности, были проанализированы три корпуса текстов:
  - Статьи из базы данных PubMed Central. Количество документов: 9219
  - Статьи из онлайн-библиотеки LibGen. Количество документов: 6909
  - Аннотации статей в закрытом доступе. Количество документов: 6708

  В результате анализа было найдено 13084 отношений: 8818 отношений между болезнью и бактерией и 4266 отношений между бактерией и нутриентом.
- Результаты анализа были провалидированы на размеченных вручную данных. Точность разработанного алгоритма составила 0,83 и 0,84 для отношений “бактерия-болезнь” и “бактерия-нутриент” соответственно.

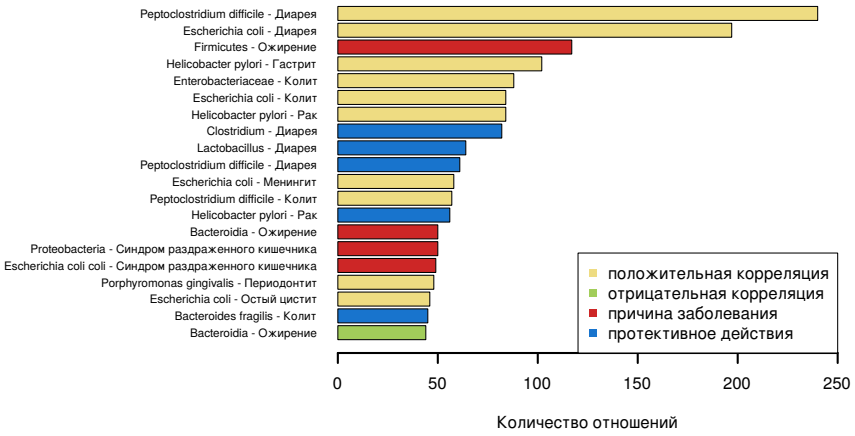

Рис. 7. Самые частоупоминаемые отношения типа “бактерия-болезнь”.

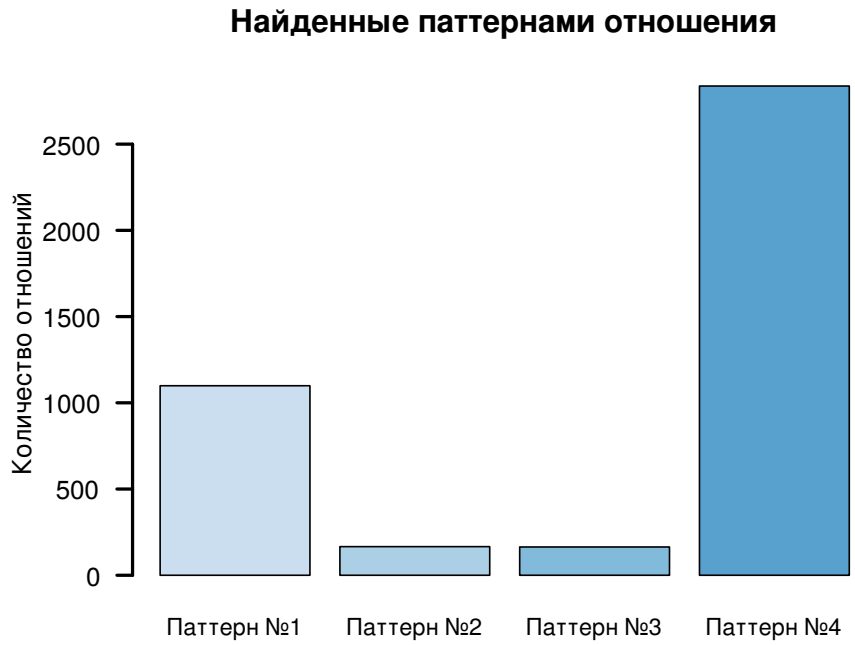

Рис. 8. Количество взаимодействий типа “бактерия-нутриент”, обнаруженных с помощью каждого паттерна.

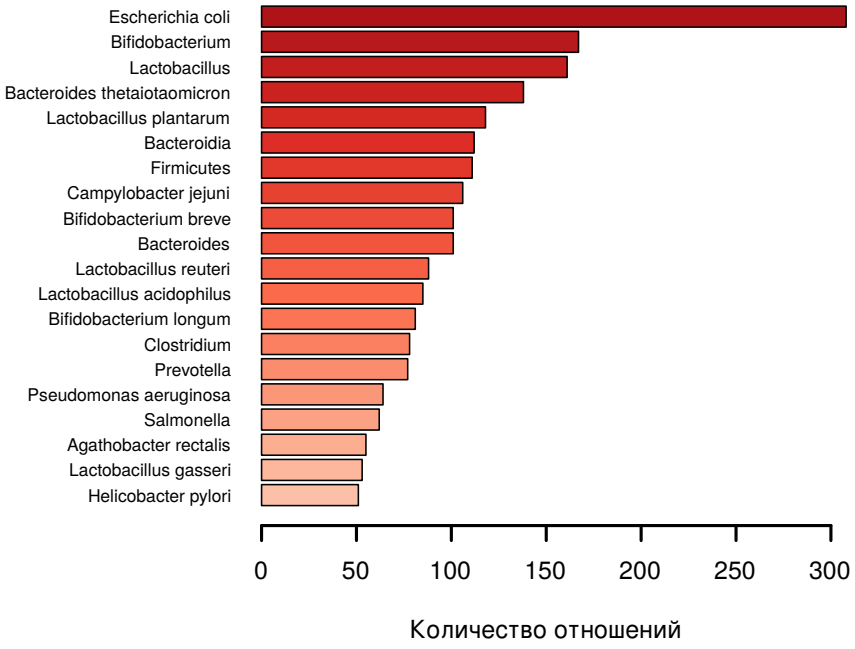

Рис. 9. Бактерии, наиболее часто входящие в отношения типа “бактерия-нутриент”.

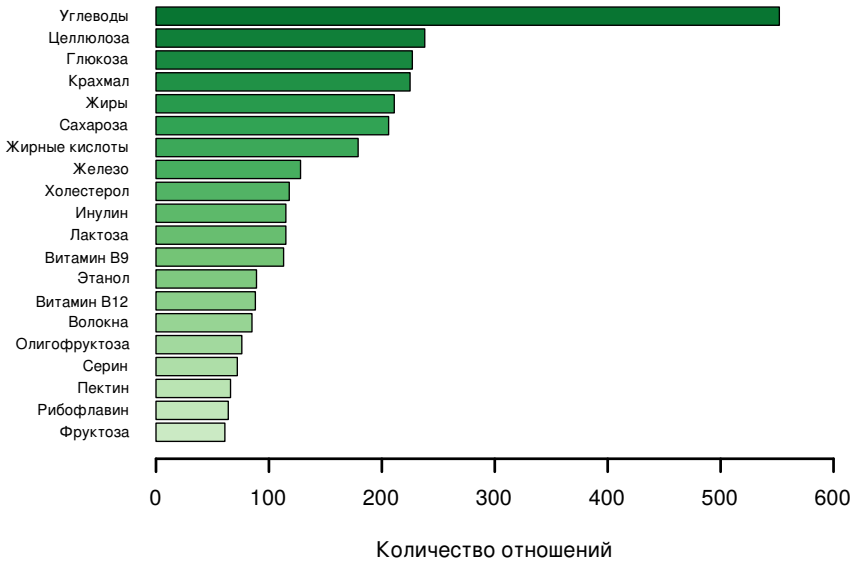

Рис. 10. Нутриеты, наиболее часто входящие в отношения типа “бактерия-нутриент”.

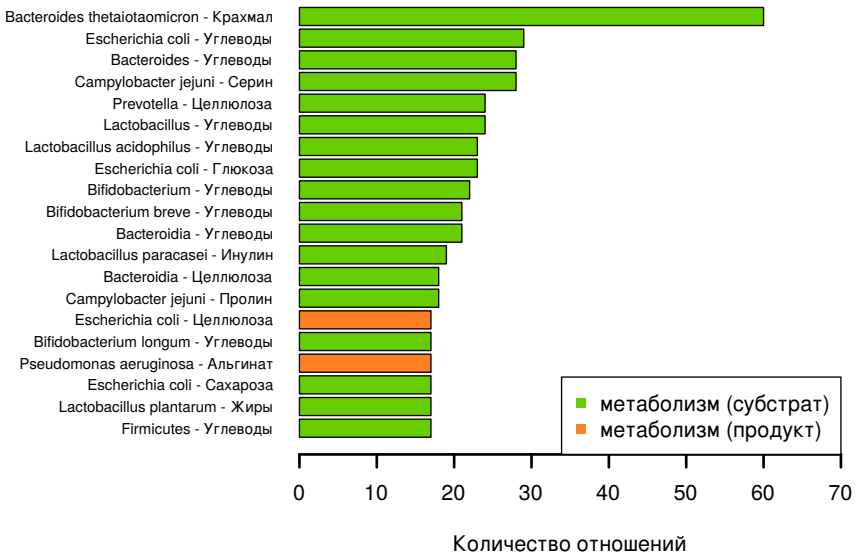

Рис. 11. Самые частоупоминаемые отношения типа “бактерия-нутриент”.

Список литературы

1. He M., Wang Y., Li W. PPI finder: a mining tool for human protein-protein interactions // PloS one. 2009. Vol. 4, no. 2. P. e4554.
2. Šarić J., Jensen L. J., Ouzounova R. et al. Extraction of regulatory gene/protein networks from Medline // Bioinformatics. 2006. Vol. 22, no. 6. P. 645–650.
3. Chen E. S., Hripcsak G., Xu H. et al. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study // Journal of the American Medical Informatics Association. 2008. Vol. 15, no. 1. P. 87–98.
4. Sender R., Fuchs S., Milo R. Revised estimates for the number of human and bacteria cells in the body // BioRxiv. 2016. P. 036103.
5. Ley R. E., Peterson D. A., Gordon J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine // Cell. 2006. Vol. 124, no. 4. P. 837–848.
6. Gebbers J., Laissue J. Immunologic structures and functions of the gut. // Schweizer Archiv fur Tierheilkunde. 1988. Vol. 131, no. 5. P. 221–238.
7. Wallace T. C., Guarner F., Madsen K. et al. Human gut microbiota and its relationship to health and disease // Nutrition Reviews. 2011. Vol. 69, no. 7. P. 392–403.
8. Renz H., Brandtzaeg P., Hornef M. The impact of perinatal immune development on mucosal homeostasis and chronic inflammation // Nature Reviews Immunology. 2012. Vol. 12, no. 1. P. 9–23.
9. Stecher B., Hardt W.-D. Mechanisms controlling pathogen colonization of the gut // Current opinion in microbiology. 2011. Vol. 14, no. 1. P. 82–91.
10. Smith K., McCoy K. D., Macpherson A. J. Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota // Seminars in immunology / Elsevier. Vol. 19. 2007. P. 59–69.
11. He T., Priebe M., Zhong Y. et al. Effects of yogurt and bifidobacteria supplementation on the colonic microbiota in lactose-intolerant subjects // Journal of Applied Microbiology. 2008. Vol. 104, no. 2. P. 595–604.
12. Liévin-Le Moal V., Servin A. L. Anti-infective activities of lactobacillus strains in the human intestinal microbiota: from probiotics to gastrointestinal anti-infectious biotherapeutic agents // Clinical microbiology reviews. 2014. Vol. 27, no. 2. P. 167–199.
13. Brigidi P., Vitali B., Swennen E. et al. Effects of probiotic administration upon

the composition and enzymatic activity of human fecal microbiota in patients with irritable bowel syndrome or functional diarrhea // Research in Microbiology. 2001. Vol. 152, no. 8. P. 735–741.

14. Gill C., Rowland I. Diet and cancer: assessing the risk // British journal of nutrition. 2002. Vol. 88, no. S1. P. s73–s87.
15. Swidsinski A., Loening-Baucke V., Verstraelen H. et al. Biostructure of fecal microbiota in healthy subjects and patients with chronic idiopathic diarrhea // Gastroenterology. 2008. Vol. 135, no. 2. P. 568–579.
16. Longstreth G. F., Thompson W. G., Chey W. D. et al. Functional bowel disorders // Gastroenterology. 2006. Vol. 130, no. 5. P. 1480–1491.
17. Norman K., Pirlich M. Gastrointestinal tract in liver disease: which organ is sick? // Current Opinion in Clinical Nutrition & Metabolic Care. 2008. Vol. 11, no. 5. P. 613–619.
18. Kim S. W., Peck K. R., Jung S. I. et al. Pseudomonas aeruginosa as a potential cause of antibiotic-associated diarrhea. // Journal of Korean medical science. 2001. Vol. 16, no. 6. P. 742.
19. Kelly C. P., Pothoulakis C., LaMont J. T. Clostridium difficile colitis // New England Journal of Medicine. 1994. Vol. 330, no. 4. P. 257–262.
20. Zoetendal E. G., Akkermans A. D., De Vos W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria // Applied and environmental microbiology. 1998. Vol. 64, no. 10. P. 3854–3859.
21. Costello E. K., Lauber C. L., Hamady M. et al. Bacterial community variation in human body habitats across space and time // Science. 2009. Vol. 326, no. 5960. P. 1694–1697.
22. Walker A. W., Ince J., Duncan S. H. et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota // The ISME journal. 2011. Vol. 5, no. 2. P. 220–230.
23. Martínez I., Kim J., Duffy P. R. et al. Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects // PloS one. 2010. Vol. 5, no. 11. P. e15046.
24. Russell W. R., Gratz S. W., Duncan S. H. et al. High-protein, reduced-carbo-

hydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health // The American journal of clinical nutrition. 2011. Vol. 93, no. 5. P. 1062–1072.

25. Flint H. J., Duncan S. H., Scott K. P., Louis P. Interactions and competition within the microbial community of the human colon: links between diet and health // Environmental microbiology. 2007. Vol. 9, no. 5. P. 1101–1111.
26. Scott K., Duncan S., Flint H. Dietary fibre and the gut microbiota // Nutrition Bulletin. 2008. Vol. 33, no. 3. P. 201–211.
27. Russell W. R., Scobbie L., Chesson A. et al. Anti-inflammatory implications of the microbial transformation of dietary phenolic compounds // Nutrition and cancer. 2008. Vol. 60, no. 5. P. 636–642.
28. Derrien M., Vaughan E. E., Plugge C. M., de Vos W. M. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium // International journal of systematic and evolutionary microbiology. 2004. Vol. 54, no. 5. P. 1469–1476.
29. De Palma G., Nadal I., Medina M. et al. Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children // BMC microbiology. 2010. Vol. 10, no. 1. P. 1.
30. Larsen N., Vogensen F. K., van den Berg F. W. et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults // PloS one. 2010. Vol. 5, no. 2. P. e9085.
31. Turnbaugh P. J., Hamady M., Yatsunenko T. et al. A core gut microbiome in obese and lean twins // nature. 2009. Vol. 457, no. 7228. P. 480–484.
32. Tannock G. Molecular analysis of the intestinal microflora in IBD // Mucosal immunology. 2008. Vol. 1. P. S15–S18.
33. De La Cochetière M.-F., Durand T., Lalande V. et al. Effect of antibiotic therapy on human fecal microbiota and the relation to the development of Clostridium difficile // Microbial ecology. 2008. Vol. 56, no. 3. P. 395–402.
34. Löfmark S., Jernberg C., Jansson J. K., Edlund C. Clindamycin-induced enrichment and long-term persistence of resistant Bacteroides spp. and resistance genes // Journal of Antimicrobial Chemotherapy. 2006. Vol. 58, no. 6. P. 1160–1167.
35. Tanaka S., Kobayashi T., Songjinda P. et al. Influence of antibiotic exposure in

the early postnatal period on the development of intestinal microbiota // FEMS Immunology & Medical Microbiology. 2009. Vol. 56, no. 1. P. 80–87.

36. McFarland L. V. Antibiotic-associated diarrhea: epidemiology, trends and treatment. 2008.
37. Pham M., Lemberg D. A., Day A. S. Probiotics: sorting the evidence from the myths // Medical Journal of Australia. 2008. Vol. 189, no. 3. P. 182–182.
38. Lupp C., Robertson M. L., Wickham M. E. et al. Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae // Cell host & microbe. 2007. Vol. 2, no. 2. P. 119–129.
39. Stecher B., Robbiani R., Walker A. W. et al. Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota // PLoS Biol. 2007. Vol. 5, no. 10. P. e244.
40. Whary M., Danon S., Feng Y. et al. Rapid onset of ulcerative typhlocolitis in B6. 129P2-IL10tm1Cgn (IL-10-/-) mice infected with Helicobacter trogontum is associated with decreased colonization by altered Schaedler's flora // Infection and immunity. 2006. Vol. 74, no. 12. P. 6615–6623.
41. Hunter L., Cohen K. B. Biomedical language processing: what's beyond PubMed? // Molecular cell. 2006. Vol. 21, no. 5. P. 589–594.
42. Zhou D., He Y. Extracting interactions between proteins from the literature // Journal of biomedical informatics. 2008. Vol. 41, no. 2. P. 393–407.
43. Jamieson D. G., Gerner M., Sarafraz F. et al. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database // Database. 2012. Vol. 2012. P. bas023.
44. Srinivasan P., Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases // Bioinformatics. 2004. Vol. 20, no. suppl 1. P. i290–i296.
45. Draghici S., Khatri P., Martins R. P. et al. Global functional profiling of gene expression // Genomics. 2003. Vol. 81, no. 2. P. 98–104.
46. Delfs R., Doms A., Kozlenkov A., Schroeder M. GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed. // German Conference on Bioinformatics / Citeseer. Vol. 169. 2004. P. 178.
47. Doms A., Schroeder M. GoPubMed: exploring PubMed with the gene ontology // Nucleic acids research. 2005. Vol. 33, no. suppl 2. P. W783–W786.

48. Dietze H., Schroeder M. GoWeb: a semantic search engine for the life science web // BMC bioinformatics. 2009. Vol. 10, no. 10. P. 1.
49. Müller H.-M., Kenny E. E., Sternberg P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature // PLoS Biol. 2004. Vol. 2, no. 11. P. e309.
50. Hoffmann R., Valencia A. Implementing the iHOP concept for navigation of biomedical literature // Bioinformatics. 2005. Vol. 21, no. suppl 2. P. ii252–ii258.
51. Munkhdalai T., Li M., Kim T. et al. Bio named entity recognition based on co-training algorithm // Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on / IEEE. 2012. P. 857–862.
52. Fukuda K.-i., Tsunoda T., Tamura A. et al. Toward information extraction: identifying protein names from biological papers // Pac symp biocomput / Citeseer. Vol. 707. 1998. P. 707–718.
53. Hanisch D., Fundel K., Mevissen H.-T. et al. ProMiner: rule-based protein and gene entity recognition // BMC bioinformatics. 2005. Vol. 6, no. Suppl 1. P. S14.
54. Tsuruoka Y., Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition // Journal of biomedical informatics. 2004. Vol. 37, no. 6. P. 461–470.
55. Tsuruoka Y., Tsujii J. Boosting precision and recall of dictionary-based protein name recognition // Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13 / Association for Computational Linguistics. 2003. P. 41–48.
56. Cohen W. W., Sarawagi S. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. 2004. P. 89–98.
57. Krauthammer M., Rzhetsky A., Morozov P., Friedman C. Using BLAST for identifying gene and protein names in journal articles // Gene. 2000. Vol. 259, no. 1. P. 245–252.
58. Ju Z., Wang J., Zhu F. Named entity recognition from biomedical text using SVM // Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on / IEEE. 2011. P. 1–4.

59. Mitsumori T., Fation S., Murata M. et al. Gene/protein name recognition based on support vector machine using dictionary as features // BMC bioinformatics. 2005. Vol. 6, no. 1. P. 1.
60. Finkel J., Dingare S., Nguyen H. et al. Exploiting context for biomedical entity recognition: from syntax to the web // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications / Association for Computational Linguistics. 2004. P. 88–91.
61. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications / Association for Computational Linguistics. 2004. P. 104–107.
62. Yang L., Zhou Y. Two-phase biomedical named entity recognition based on semi-CRFs // Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on / IEEE. 2010. P. 1061–1065.
63. Chan S.-K., Lam W. Efficient Methods for Biomedical Named Entity Recognition // 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering / IEEE. 2007. P. 729–735.
64. Dredze M., McNamee P., Rao D. et al. Entity disambiguation for knowledge base population // Proceedings of the 23rd International Conference on Computational Linguistics / Association for Computational Linguistics. 2010. P. 277–285.
65. Bunescu R. C., Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation. // EACL. Vol. 6. 2006. P. 9–16.
66. Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. // EMNLP-CoNLL. Vol. 7. 2007. P. 708–716.
67. Aggarwal C. C., Zhai C. Mining text data. Springer Science & Business Media, 2012.
68. Zhou D., Zhong D., He Y. Biomedical relation extraction: from binary to complex // Computational and mathematical methods in medicine. 2014. Vol. 2014.
69. Proux D., Rechenmann F., Julliard L. et al. A pragmatic information extraction strategy for gathering data on genetic interactions. // ISMB. Vol. 8. 2000. P. 279–285.
70. Raja K., Subramani S., Natarajan J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature // Database. 2013. Vol. 2013. P. bas052.

71. Rastier F., Cavazza M., Abeillé A. Semantics for descriptions: From linguistics to computer science. No. 138. Stanford Univ Center for the Study, 2002.
72. Clegg A. B., Shepherd A. J. Benchmarking natural-language parsers for biological applications using dependency graphs // BMC bioinformatics. 2007. Vol. 8, no. 1. P. 1.
73. Fundel K., Küffner R., Zimmer R. RelEx—Relation extraction using dependency parse trees // Bioinformatics. 2007. Vol. 23, no. 3. P. 365–371.
74. Björne J., Heimonen J., Ginter F. et al. Extracting complex biological events with rich graph-based feature sets // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task / Association for Computational Linguistics. 2009. P. 10–18.
75. Björne J., Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task // Proceedings of the BioNLP Shared Task 2013 Workshop. 2013. P. 16–25.
76. Kilicoglu H., Bergler S. Syntactic dependency based heuristics for biological event extraction // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task / Association for Computational Linguistics. 2009. P. 119–127.
77. Bach N., Badaskar S. A review of relation extraction // Literature review for Language and Statistics II. 2007.
78. Miwa M., Sætre R., Miyao Y., Tsujii J. A rich feature vector for protein-protein interaction extraction from multiple corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 / Association for Computational Linguistics. 2009. P. 121–130.
79. Song M., Kim W. C., Lee D. et al. PKDE4J: Entity and relation extraction for public knowledge discovery // Journal of biomedical informatics. 2015. Vol. 57. P. 320–332.
80. Miwa M., Sætre R., Kim J.-D., Tsujii J. Event extraction with complex event classification using rich features // Journal of bioinformatics and computational biology. 2010. Vol. 8, no. 01. P. 131–146.
81. Miwa M., Thompson P., Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution // Bioinformatics. 2012. Vol. 28, no. 13. P. 1759–1765.

82. Lodhi H., Saunders C., Shawe-Taylor J. et al. Text classification using string kernels // Journal of Machine Learning Research. 2002. Vol. 2, no. Feb. P. 419–444.
83. Mooney R. J., Bunescu R. C. Subsequence kernels for relation extraction // Advances in neural information processing systems. 2005. P. 171–178.
84. Moschitti A. Making Tree Kernels Practical for Natural Language Learning. // EACL. Vol. 113. 2006. P. 24.
85. Airola A., Pyysalo S., Björne J. et al. A graph kernel for protein-protein interaction extraction // Proceedings of the workshop on current trends in biomedical natural language processing / Association for Computational Linguistics. 2008. P. 1–9.
86. Bunescu R. C., Mooney R. J. A shortest path dependency kernel for relation extraction // Proceedings of the conference on human language technology and empirical methods in natural language processing / Association for Computational Linguistics. 2005. P. 724–731.
87. Miwa M., Sætre R., Miyao Y. et al. Combining multiple layers of syntactic information for protein-protein interaction extraction // Proceedings of the third international symposium on semantic mining in biomedicine. 2008. P. 101–108.
88. Bravo À., Piñero J., Queralt-Rosinach N. et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research // BMC bioinformatics. 2015. Vol. 16, no. 1. P. 1.
89. Kim J.-D., Ohta T., Tateisi Y., Tsujii J. GENIA corpus—a semantically annotated corpus for bio-textmining // Bioinformatics. 2003. Vol. 19, no. suppl 1. P. i180–i182.
90. Hirschman L., Yeh A., Blaschke C., Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology // BMC bioinformatics. 2005. Vol. 6, no. 1. P. 1.
91. Krallinger M., Morgan A., Smith L. et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge // Genome biology. 2008. Vol. 9, no. 2. P. 1.
92. Leitner F., Mardis S. A., Krallinger M. et al. An overview of BioCreative II. 5 // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010. Vol. 7, no. 3. P. 385–399.
93. Arighi C. N., Lu Z., Krallinger M. et al. Overview of the BioCreative III workshop // BMC bioinformatics. 2011. Vol. 12, no. 8. P. 1.

94. Arighi C. N., Carterette B., Cohen K. B. et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task // Database. 2013. Vol. 2013. P. bas056.
95. Simpson M. S., Demner-Fushman D. Biomedical text mining: A survey of recent progress // Mining text data. Springer, 2012. P. 465–517.
96. Névéol A., Lu Z. Automatic integration of drug indications from multiple health resources // Proceedings of the 1st ACM international health informatics symposium / ACM. 2010. P. 666–673.
97. Srinivasan P., Rindflesch T. Exploring text mining from MEDLINE. // Proceedings of the AMIA Symposium / American Medical Informatics Association. 2002. P. 722.
98. Wang X., Hripcsak G., Markatou M., Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study // Journal of the American Medical Informatics Association. 2009. Vol. 16, no. 3. P. 328–337.
99. Swanson D. R. Complementary structures in disjoint science literatures // Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval / ACM. 1991. P. 280–289.
100. Swanson D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge // Perspectives in biology and medicine. 1986. Vol. 30, no. 1. P. 7–18.
101. Ramadan N., Halvorson H., Vande-Linde A. et al. Low brain magnesium in migraine // Headache: The Journal of Head and Face Pain. 1989. Vol. 29, no. 9. P. 590–593.
102. Ferrari M. Biochemistry of migraine. // Pathologie-biologie. 1992. Vol. 40, no. 4. P. 287–292.
103. Weeber M., Vos R., Klein H. et al. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide // Journal of the American Medical Informatics Association. 2003. Vol. 10, no. 3. P. 252–259.
104. Srinivasan P., Libbus B., Sehgal A. K. Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases // Workshop BioLINK, linking biological literature, ontologies and databases at HLT NAACL. 2004. P. 33–40.
105. Frijters R., Van Vugt M., Smeets R. et al. Literature mining for the discovery of

hidden connections between drugs, genes and diseases // PLoS Comput Biol. 2010. Vol. 6, no. 9. P. e1000943.

106. Jelier R., Schuemie M. J., Veldhoven A. et al. Anni 2.0: a multipurpose text-mining tool for the life sciences // Genome biology. 2008. Vol. 9, no. 6. P. 1.
107. Yang Y., Adelstein S. J., Kassis A. I. Target discovery from data mining approaches // Drug discovery today. 2009. Vol. 14, no. 3. P. 147–154.
108. van Haagen H. H., AC't Hoen P., Bovo A. B. et al. Novel protein-protein interactions inferred from literature context // PLoS One. 2009. Vol. 4, no. 11. P. e7894.
109. Elkin P. L., Tuttle M. S., Trusko B. E., Brown S. H. BioProspecting: novel marker discovery obtained by mining the bibleome // BMC bioinformatics. 2009. Vol. 10, no. 2. P. 1.
110. Yang Z., Lin H., Li Y. BioPPISVMExtractor: A protein–protein interaction extractor for biomedical literature using SVM and rich feature sets // Journal of biomedical informatics. 2010. Vol. 43, no. 1. P. 88–96.
111. Narayanaswamy M., Ravikumar K., Vijay-Shanker K. Beyond the clause: extraction of phosphorylation information from medline abstracts // Bioinformatics. 2005. Vol. 21, no. suppl 1. P. i319–i327.
112. Özgür A., Xiang Z., Radev D. R., He Y. Literature-Based Discovery of IFN-gamma and Vaccine-Mediated Gene Interaction Networks // BioMed Research International. 2010. Vol. 2010.
113. Jelier R., Jenster G., Dorssers L. C. et al. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation // BMC bioinformatics. 2007. Vol. 8, no. 1. P. 1.
114. Ritari J., Salojärvi J., Lahti L., de Vos W. M. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database // BMC genomics. 2015. Vol. 16, no. 1. P. 1056.
115. Schriml L. M., Arze C., Nadendla S. et al. Disease Ontology: a backbone for disease semantic integration // Nucleic acids research. 2012. Vol. 40, no. D1. P. D940–D946.
116. Bird S., Klein E., Loper E. Natural language processing with Python. "O'Reilly Media, Inc. 2009.
117. Chen D., Manning C. D. A Fast and Accurate Dependency Parser using Neural Networks. // EMNLP. 2014. P. 740–750.