Московский физико-технический институт (государственный университет)
Факультет биологической и медицинской физики
Кафедра Молекулярной 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 биотехнологи

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 3
2. Введение 4
3. Литературный обзор 6

1. Список сокращений

ЖКТ — Желудочно-кишечный тракт

ИАТ — Интеллектуальный анализ текстов

ИАБТ — Интеллектуальный анализ биомедицинских текстов

КЖК — Короткоцепочечные жирные кислоты

МЕММ, 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 [**?**], ген-белковых взаимодействий [**?**], определение мишеней лекарственных препаратов [**?**] 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