АННОТАЦИЯ

на диссертационную работу Голенко Екатерины Сергеевны на тему «Разработка алгоритмов анализа данных масс-спектрометрии нативных белков» на соискание степени доктора философии (PhD) по образовательной программе 8D06101 – «Аналитика больших данных»

Актуальность темы исследования. В современной научной практике масс-спектрометрия утвердилась как одна их центральных технологий для анализа пептидов и белков. Процедура идентификации белков с помощью масс-спектрометрии включает расщепление белков на пептиды, которые затем разделяются, фрагментируются, ионизируются и улавливаются масс-спектрометрами. Белки распознаются и каталогизируются на основе их масс-спектров, точнее, по характерным пикам, которые соответствуют ионам пептидных фрагментов. Множество факторов, включая присутствие посттрансляционных модификаций, разложение белков на фрагменты и проблемы с ионизацией, влияют на то, что только ограниченный набор белков может быть точно идентифицирован. Современные методы обычно позволяют достоверно идентифицировать менее половины всех белков в образце, оставляя значительное количество потенциально важных молекул без подтверждения их присутствия и функции. Это требует дальнейшего усовершенствования методов масс-спектрометрии и разработки более мощных алгоритмов для обработки и интерпретации спектральных данных.

Текущие стратегии идентификации белков в биоинформатике делятся на две основные категории. Первая – это поиск по базам данных, функционирующий на принципе сопоставления экспериментально полученных масс-спектров с пулом теоретически сгенерированных спектров пептидов, созданных *in silico* на основании известных и секвенированных последовательностей белков. Обычно используются базы данных белков с поисковыми системами, например, Mascot, SEQUEST и X!Tandem. Поиск по спектральным библиотекам представляет собой передовую методологию, использующую библиотеки, включающие данные о масс-спектрах пептидов, полученные в ходе предшествующих экспериментов. Когда появляется масс-спектр неизвестного пептида, его сравнивают с имеющимися спектрами в библиотеке для точной идентификации, повышая скорость анализа и улучшая точность определения пептидов. Наконец, метод секвенирования *de novo* определяет последовательности белков напрямую из масс-спектрометрических данных, исследуя пики, соответствующие фрагментированным пептидным ионам, без применения биологических баз данных.

Использование баз данных для идентификации пептидов и белков является предпочтительным методом анализа, этот подход эффективен для распознавания уже изученных белков, чьи последовательности содержатся в базах данных, но ограничен при поиске неизвестных белков или тех, которые претерпели посттрансляционные изменения, что может увеличить время поиска и риск получения неверных результатов.

Прогнозирование функций белка является ещё одной серьёзной задачей биоинформатики, целью которой является определение функций, выполняемых известным белком. Как правило, идентификация функций белка осуществляется посредством ручной или компьютерной аннотации. Автоматизированное прогнозирование функций на основе системы Gene Ontology является сложнейшей задачей биоинформатики. Сложность этой задачи обусловлена следующими факторами: во-первых, большая часть белков, не подвергшихся аннотированию экспертами, не содержит никакой информации кроме аминокислотной последовательности. Во-вторых, возникает сложный вопрос настройки параметров и гиперпараметров выбранной модели. В-третьих, Gene Ontology имеет сложную и неоднородную структуру, поэтому задача прогнозирования функция должна рассматриваться как задача со множественными выходными метками.

Таким образом, сегодня значимость представляют алгоритмы следующих направленностей: (i) алгоритмы, фокусирующиеся на первичной обработке информации, что облегчает трансформацию сырых спектрометрических данных для более глубокого анализа; (ii) алгоритмы интерпретации спектральных данных для выявления общих и неявных закономерностей; (iii) базы данных и связанные с ними алгоритмы для идентификации пептидов и белков, и (iv) алгоритмы функциональной аннотации, способные ассоциировать белковые последовательности с биологическими функциями.

Целью диссертации является разработка алгоритмов для интерпретации результатов масс-спектрометрии и предсказания функций белков.

Задачи диссертационного исследования:

1. Провести анализ существующих решений для обработки данных масс-спектрометрии и белковых последовательностей;
2. Разработать алгоритм для идентификации пептидов с применением моделей машинного обучения;
3. Разработать алгоритм для предсказания функций белковых последовательностей с помощью методов машинного обучения;
4. Провести оценку предложенных алгоритмов на общедоступных наборах данных с использованием общепринятых для машинного обучения оценочных показателей.

Объектом диссертационного исследования являются данные масс-спектрометрии, полученные в рамках проекта «ПЦР-тест для детекции и дифференциальной диагностики возбудителей описторхоза и меторхоза» под руководством Киян В. С., PhD, НИПСБ, а также данные из общедоступных баз данных белковых и ДНК-последовательностей, в том числе NIST, PRIDE, Pfam, GenBank и UniProtKB/Swiss-Prot.

Предметом диссертационного исследования являются алгоритмы для идентификации пептидов и белков, а также определения их функций.

Методы исследования. При анализе экспериментальных данных и разработке алгоритмов были использованы методы анализа больших массивов данных, качественный анализ, методы сравнения последовательностей, нейронные сети, кластеризация и классификация.

Основные научные положения, выносимые на защиту и обладающие признаками научной новизны:

1. Алгоритм для идентификации пептидов, полученных путём масс-спектрометрии, основанный на двунаправленной нейронной сети LSTM, заложенной в сети глубокого подобия для работы со спектрами и пептидами;

2. Алгоритм для предсказания функций белковых последовательностей, основанный на двунаправленной нейронной сети LSTM и механизме «self-attention».

Связь темы с планами научно-исследовательских программ. Тема исследования соответствует приоритетному научному направлению «Информационные, коммуникационные и космические технологии».

Представленные результаты получены при выполнении проекта AP05131132 «ПЦР-тест для детекции и дифференциальной диагностики возбудителей описторхоза и меторхоза» в 2018-2020 годы.

Научная новизна диссертационного исследования:

- Предложен алгоритм для идентификации пептидов, разработанный на основе сети подобия с открытым исходным кодом SpeCollate, использующий нейронную сеть BiLSTM для поиска совпадений пептидного спектра.
- Предложен алгоритм аннотации белковых функций, построенный на основе комбинации нейронной сети BiLSTM и механизма самовнимания («self-attention»).
- Разработанные алгоритмы позволяют обрабатывать биологические данные, одновременно используя как машинное обучение, так и методы сравнения последовательностей.

Теоретическая значимость результатов диссертационного исследования:

- Разработанные алгоритмы не имеют ограничений на длину аминокислотной последовательности и, следовательно, могут использоваться для аннотации функций белка в масштабе генома.
- Работают быстро и могут аннотировать несколько тысяч белков за несколько минут даже на одном процессоре.
- Модели не ограничены несбалансированной или отсутствующей информацией о межбелковых взаимодействиях.
- Алгоритмы можно применять, используя только мотивы последовательности.

Практическая значимость результатов диссертационного исследования:

- Разработанные алгоритмы могут быть внедрены программные модули лабораторий для идентификации белковых последовательностей и предсказания их функций с высокой надёжностью.
- Разработанные алгоритмы могут быть использованы биологами как альтернативный существующим приложениям либо дополнительный инструментарий при работе с биологическими данными.
- Заложенные в основе алгоритмов нейронные сети могут обучаться на различных наборах данных для решения широкого круга биологических

проблем, требующих определения функций белков, в первую очередь для понимания механизмов болезней и разработки лекарств.

Научно-обоснованные теоретические и экспериментальные результаты диссертационной работы использованы в научном проекте по теме «ПЦР-тест для детекции и дифференциальной диагностики возбудителей описторхоза и меторхоза».

Созданные в результате диссертационного исследования программные модули нашли применение и внедрены в лаборатории биоразнообразия и генетических ресурсов «Национального центра биотехнологии» (Астана, Казахстан) и ООО «Новые программные системы» (Новосибирск, РФ).

Апробация результатов диссертационного исследования. Основные результаты диссертационного исследования докладывались и обсуждались на научных семинарах кафедры «Информационные системы» КАТУ им С. Сейфуллина, кафедры «Информационные системы» ЕНУ им. Л. Н. Гумилёва и на следующих международных научно-практических конференциях:

- Международная научно-теоретическая конференция «Сейфуллинские чтения – 16», КАТИУ им. С. Сейфуллина, Нур-Султан, 2020 г.;
- Международная научно-практическая конференция «Интеграция науки, образования и производства основа реализации Плана Нации», КарГТУ, 2020 г.;
- Международная научно-теоретическая конференция «Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация»», Нур-Султан, 2021 г.;
- Международная научная конференция «XXII Сатпаевские чтения», Satbayev University, Алматы, 2022 г.;
- Международная научно-теоретическая конференция «Сейфуллинские чтения – 18: «Молодёжь и наука – взгляд в будущее»», КАТИУ им. С. Сейфуллина, Астана, 12 апреля 2022 г.;
- Международная научная конференция «Математическая логика и компьютерные науки», ЕНУ им. Л. Н. Гумилёва, Астана, 7-8 октября 2022 г.

Статьи в журналах, входящих в международную базу SCOPUS. Вклад автора состоит в выдвижении гипотезы, сборе данных, технической реализации экспериментов, интерпретации результатов и подготовке публикации:

1. Golenko et al. Implementation of machine learning models to determine the appropriate model for protein function prediction. Eastern-European Journal of Enterprise Technologies, 2022. https://doi.org/10.15587/1729-4061.2022.263270

Статьи в журналах, рекомендованных Комитетом по обеспечению качества в сфере науки и высшего образования Министерства науки и высшего образования Республики Казахстан. В рамках данных публикаций автор является разработчиком исследовательской концепции и методологии анализа данных. Непосредственно участвовал в сборе данных, интерпретации результатов, формулировке выводов и подготовке научных статей к публикации:

1. Голенко Е., Исмаилова А., Жумаханова А. Предсказание функций белков при помощи базы данных «Gene Ontology» и моделей машинного обучения. Известия НАН РК. Серия физико-математическая. №2 (342). – 2022. – С. 19–38.

2. Голенко Е., Исмаилова А., Молдашева Р. Применение методов глубокого обучения для предсказания структуры белков. Вестник Национальной инженерной академии РК. Серия Инф.-комм. технологии. №4 (86). – 2022. – С. 28–40.

3. Голенко Е., Исмаилова А. Предсказание функций белка с использованием комбинации BiLSTM и алгоритма самовнимания. Известия НАН РК. Серия физико-математическая. №3 (347). – 2023. – С. 62–75.

Личный вклад автора состоит в непосредственном выполнении исследований по всем главам и логическим звеньям диссертации: проведение обзора и анализа ранее представленных работ, выбор и обоснование использованных методов, разработка и техническая реализация алгоритмов, апробация и тестирование разработанных моделей на исходных данных.

Публикации по теме диссертационного исследования. По теме диссертационного исследования было опубликовано 13 (тринадцать) научных трудов, из них 1 (одна) статья в научном журнале с ненулевым импакт-фактором, входящим в международную базу SCOPUS (процентиль по CiteScore2022 равный 34), 3 (три) статьи в журналах, рекомендованных Комитетом по обеспечению качества в сфере науки и высшего образования Министерства науки и высшего образования Республики Казахстан, 6 (шесть) статей в сборниках международных конференций, 3 (три) статьи – в других изданиях. Имеется 2 (два) авторских свидетельства о государственной регистрации программы для ЭВМ.

Структура и объём диссертационной работы. Диссертационное исследование представлено в следующем формате: введение, три основных раздела, заключение, список использованных источников (123 наименования) и два приложения. Общий объём составляет 123 страницы компьютерного текста, сопровождается 23 рисунками и 7 таблицами.

Во введении подчёркивается важность изучаемой темы, уровень изученности, раскрыта актуальность разработанных алгоритмов для обработки данных протеомики, сформулирована цель исследования, поставлены задачи, определены предмет и объект исследования, раскрыты научная новизна, теоретическая и практическая значимость исследования. Приводятся данные об апробации и публикациях результатов исследования, а также указывается личный вклад автора в научные исследования.

В первом разделе проанализировано текущее состояние международных глобальных репозиториев белковых структур и генетических последовательностей, из которого вытекает постановка двух главных задач диссертационного исследования. Проведён обширный анализ проблемы идентификации белков и пептидов, выделенных посредством масс-спектрометрического анализа, а также изучены методы оценки корректности идентифицированных пептидов. Рассмотрены методы для решения задачи идентификации белков и пептидов. Помимо этого, осуществлён анализ проблематики аннотации белков, полученных экспериментальным путём, и обзор алгоритмов для функционального прогнозирования. Выявлены ключевые недостатки и направления для усовершенствования этих алгоритмов.

Во втором разделе предложено решение для идентификации пептидов и белков на основе общедоступной сети SpeCollate, адаптированной для создания вложений спектров и пептидов в единое евклидово пространство. Охарактеризован процесс обучения сети на данных, представляющих как положительные, так и отрицательные примеры, и осуществляется в контексте с функцией потерь SNAP, способствующей эффективному различению между соответствующими и несоответствующими парами. Представлены результаты обучения модели и проведена оценка её эффективности.

В третьем разделе представлен процесс разработки алгоритма для функционального аннотирования белковых последовательностей. Описан процесс предварительной обработки и анализа экспериментальных данных, полученных из открытых источников, для обучения нейронной сети. Представлена модель алгоритма с использованием двунаправленной LSTM в комбинации с механизмом «self-attention». Приведены результаты обучения модели на экспериментальных данных. Проведена оценка надёжности предсказания функций разработанной модели. Также приведены результаты ручного аннотирования функций белка.

В заключении представлены результаты исследования, сформулированы основные выводы, подтверждающие и доказывающие истинность положений, выносимых на защиту.

В приложениях представлены авторские свидетельства и акты внедрения.