Казахский агротехнический исследовательский университет
им. С. Сейфуллина

УДК 004.8:57.08 (043) На правах рукописи

ГОЛЕНКО ЕКАТЕРИНА СЕРГЕЕВНА

Разработка алгоритмов анализа данных масс-спектрометрии нативных белков

8D06101 – Аналитика больших данных

Диссертация на соискание степени
доктора философии (PhD)

Научный консультант:
доктор PhD,
ассоциированный профессор
Исмаилова А. А.

Зарубежный научный консультант:
кандидат физико-математических наук,
старший научный сотрудник
Штокало Д. Н.
(ИСИ СО РАН, Новосибирск, РФ)

Республика Казахстан
Астана, 2024

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ

В настоящей диссертации используются следующие термины с соответствующими определениями:

Масс-спектрометрия – это аналитический метод, используемый для измерения отношения массы к заряду ионов *m/z*. Результаты представляются в виде масс-спектра, графика интенсивности в функции отношения массы к заряду. Этот метод позволяет определять элементарный или изотопный состав образца, массы частиц и молекул, а также анализировать химическую структуру молекул и других химических соединений.

Протеомика – это раздел молекулярной биологии, отвечающий за идентификацию и количественный анализ белков и пептидов.

Белки – это большие, сложные молекулы, которые выполняют множество критически важных функций в живых организмах. Они состоят из одной или нескольких длинных цепочек аминокислот и необходимы для структуры, функции и регуляции тканей и органов. Белки участвуют в процессах в клетках, способствуют метаболизму за счёт катализа биохимических реакций, помогают в репликации ДНК, отвечают за реакции на стимулы и многое другое.

Пептиды – это короткие цепочки аминокислот, соединённые пептидными связями. В отличие от белков, которые обычно состоят из 50 или более аминокислот, пептиды меньше и могут содержать от двух до нескольких десятков аминокислот. Они играют множество ролей в организме, в том числе служат строительными блоками для более крупных белковых структур.

Аминокислоты – это органические соединения, состоящие из аминогруппы (-NH2), карбоксильной группы (-COOH) и боковой цепи, уникальной для каждой аминокислоты, присоединенной к центральному альфа-углероду. Существует 20 стандартных аминокислот, которые кодируются генетическим кодом и используются для синтеза белков в клетках.

Секвенирование *de novo* – метод, при котором аминокислотная последовательность пептида определяется с помощью тандемной масс-спектрометрии.

Первичная структура белка – простейший вид белковой структуры, представляющий собой последовательность аминокислот в полипептидной цепи.

Биологические базы данных – базы данных, состоящие из биологических данных, таких как последовательности белков, молекулярная структура, генетические последовательности и т. д., в организованной форме.

Gene Ontology – это всеобъемлющий современный биологический сборник, который используется в качестве ресурса для аннотации генов и генных продуктов всех биологических видов.

PFam – база данных семейств белковых доменов.

GenBank – база данных, находящаяся в открытом доступе, содержащая все аннотированные последовательности ДНК и РНК, а также последовательности закодированных в них белков.

UniProt – открытая база данных последовательностей белков.

UniProtKB/Swiss-Prot – рецензируемый экспертами раздел базы знаний UniProt.

Машинное обучение – это раздел искусственного интеллекта, который основан на идее, что системы могут учиться из данных, идентифицировать закономерности и принимать решения с минимальным участием человека. Это достигается путём создания алгоритмов, которые могут получать данные и использовать статистический анализ для прогнозирования и оптимизации результатов.

Модель машинного обучения – это математическая модель, которая использует алгоритмы для анализа данных, обучения и принятия прогностических или классификационных решений. В её основе лежат вычислительные методы, которые позволяют модели улучшать свои предсказания или решения с увеличением количества обрабатываемой информации.

Классификация – в контексте машинного обучения это тип задачи с обучением под наблюдением, при которой модель обучается распознавать, к какому классу или категории относится каждый образец в данных. Цель классификации заключается в том, чтобы после обучения на наборе данных с известными категориями (называемых обучающими данными) модель могла корректно определить класс для новых, ранее неизвестных образцов. В процессе классификации используются следующие шаги: обучение (модель обучается на основе обучающего набора данных, где каждый образец имеет метку класса), оценка (оценивается точность модели на тестовом наборе данных, чтобы увидеть, насколько хорошо она обобщает то, чему она научилась) и применение (модель используется для предсказания классов новых образцов данных).

Кластеризация – в машинном обучении это процесс группировки набора объектов таким образом, чтобы объекты в одной группе (или кластере) были более похожи (в некотором смысле) друг на друга, чем на объекты в других кластерах. Это вид обучения без учителя, так как он не требует меток классов для объектов.

Евклидово пространство – это математическое пространство, определённое в контексте евклидовой геометрии, которое характеризуется наличием понятий расстояния между точками (данное расстояние измеряется с помощью евклидовой метрики) и углов между векторами. Применительно к машинному обучению и обработке данных, например спектров, часто используется преобразование данных таким образом, чтобы они были отображены на поверхность единичной гиперсферы. Это позволяет нормализовать данные и упростить их последующий анализ, поскольку все данные будут иметь одинаковую "длину" в евклидовом пространстве, что делает их сравнимыми по их направлениям.

Функция белка – это биологическая роль, которую белок выполняет в живом организме. Это может включать каталитическую активность в ферментах, транспортировку молекул, поддержание структуры клеток и тканей, участие в сигнальных путях, регуляцию генной экспрессии и многие другие процессы.

Аннотирование белков – это процесс добавления информации о функциях, структуре, доменах, известных взаимодействиях и других биологических аспектах к белковым последовательностям. Это может включать идентификацию функциональных доменов, сайтов связывания, посттрансляционных модификаций, филогенетических связей и других функциональных данных, которые помогают исследователям понять роль белка в клетке или организме.

Посттрансляционные модификации – это химические изменения, которые происходят с белком после того, как он был синтезирован на рибосоме посредством процесса трансляции. Эти модификации могут существенно изменять физические и химические свойства белка, включая его структуру, функциональную активность, стабильность и местоположение в клетке.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ББД	– биологические базы данных
PDB	– Protein Data Bank
ENA	– European Nucleotide Archive
GO	– Gene Ontology
МС/МС	– масс-спектрометрия
EMBL	– European Molecular Biology Laboratory
SCOP	– Structural Classification of Proteins
LSTM	– Long Short-Term Memory
CNN	– Convolutional Neural Network
ДНК	– дезоксирибонуклеиновая кислота
РНК	– рибонуклеиновая кислота
SRA	– Sequence Read Archive
INSDC	– International Nucleotide Sequence Database Collaboration
DDBJ	– DNA Data Bank of Japan
NIST	– National Institute of Standards and Technology
Pfam	– Protein Families Database
PSM	– Peptide-Spectrum Matches (совпадения пептидного спектра)
FDR	– False Discovery Rates (доля ложных отклонений)
NCBI	– National Centre for Biotechnology Information
CID	– Collision Induced Dissociation (диссоциация, индуцированная столкновением)
ПТМ	– посттрансляционные модификации
OMSSA	– Open Mass Spectrometry Search Algorithm
MS-GF	– Mass-Spectrometry Generating Function (метод производящих функций),
PRM	– Prefix-Residue Mass (спектр массы префикса-остатка)
IEA	– Inferred from Electronic Annotation (полученный из электронной аннотации)
NGS	– Next-Generation Sequencing
EC	– Enzyme Commission
FunCat	– функциональный каталог
KEGG	– Киотская энциклопедия генов и геномов
МФ	– молекулярная функция
БП	– биологический процесс
КК	– клеточный компонент
DAG	– Directed Acyclic Graph (направленный ациклический граф)
PLSAnorm	– Probabilistic Latent Semantic Analysis With Normalization (вероятностный латентно-семантический анализ с нормализацией)
AFP	– Automated Function Prediction (автоматическое прогнозирование функций)

BLAST – The Basic Local Alignment Search Tool
PPI – Protein-Protein Interaction (белок-белковое взаимодействие)
MSA – Multiple Sequence Alignment (множественное выравнивание последовательностей)
PSSM – Position-Specific Scoring Matrix (матрица подсчёта очков для конкретной позиции)
MRF – Hierarchical Markov Random Field
SIM – семантически улучшенное tSVD
SVM – Support Vector Machine
KNN – k-Nearest Neighbor
LTR – Learning-To-Rank (модель обучения для ранжирования)
DNN – Deep Neural Network (глубокая нейронная сеть)
RNN – Recurrent Neural Network (рекуррентная нейронная сеть)
FCDN – Fully Connected Deep Network (полносвязная глубокая сеть)
AE – Autoencoder
RBM – Restricted Boltzmann Machine
GAN – Generative Adversarial Network
CMM – скрытая марковская модель
TSVD – Truncated Singular Value Decomposition (разложение усечённого сингулярного значения)
Да (Da) – дальтон или атомная единица массы
PSN – Peptide Sub-Network
SSN – Spectral Sub-Network

ВВЕДЕНИЕ

Проблема и актуальность темы исследования. В современной научной практике масс-спектрометрия утвердилась как одна их центральных технологий для анализа пептидов и белков, благодаря высоким показателям в аспектах скорости и точности измерений [1]. 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. Этот процесс зависит от сложных алгоритмов и вычислительных методов, способных сопоставлять экспериментально полученные масс-спектры с теоретическими массами пептидов, предсказанными из известных белковых последовательностей. Тем не менее, учитывая высокую сложность биологических образцов и ограничения современных аналитических методов, задача идентификации белков по тандемным масс-спектрам остаётся чрезвычайно трудоёмкой. 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 создаёт существенный барьер для понимания сложных биологических систем и требует дальнейшего усовершенствования методов масс-спектрометрии и разработки более мощных алгоритмов для обработки и интерпретации спектральных данных.

Также усовершенствования технологий привели к увеличению скорости обработки экспериментальных образцов, что, в свою очередь, породило большие объёмы данных, ожидающих дальнейшего анализа и обработки. Развитие программного обеспечения и вычислительных инструментов позволяет автоматизировать обработку больших объёмов данных, однако это также ставит перед исследователями задачу овладения многочисленными специализированными программными пакетами, каждый из которых имеет свои алгоритмы и интерфейсы. Поэтому сегодня наблюдается необходимость разработки новых методов идентификации белков, способных с более высокой точностью идентифицировать пептиды и белки.

Текущие стратегии идентификации белков в биоинформатике делятся на две основные категории, каждая со своим уникальным подходом и комплектом аналитических инструментов. Первый – это базы данных поисковых подходов, которые являются доминирующим и предпочтительным методом в большинстве лабораторных настроек, благодаря их эффективности и доступности. Эти методы функционируют на принципе сопоставления экспериментально полученных масс-спектров с огромным пулом теоретически сгенерированных спектров пептидов, созданных *in silico* на основании известных и секвенированных последовательностей белков, хранящихся в обширных базах

данных. Поиск сравнивает измеренные массы пептидных фрагментов с этими виртуальными базами данных, чтобы выявить наилучшие соответствия и определить идентичность белка.

Второй метод, секвенирование *de novo*, представляет собой более сложный и менее распространённый подход, который позволяет определить последовательность аминокислот непосредственно из экспериментальных масс-спектров, без предварительного сопоставления с базами данных. Этот метод полезен в случаях, когда белок отсутствует в доступных базах данных, или когда исследователи сталкиваются с новыми или ранее неизвестными белками. Секвенирование *de novo* требует значительных вычислительных мощностей и продвинутых алгоритмов для интерпретации сложных спектров и может быть подвержено ошибкам из-за посттрансляционных модификаций или других сложностей в структуре белков.

Оба метода играют ключевую роль в современной протеомике, однако, несмотря на их важность, эти методы всё ещё сталкиваются с ограничениями в точности и полноте покрытия протеома, что стимулирует непрерывные усилия в улучшении технологий масс-спектрометрии и разработке новых вычислительных решений.

В области протеомики для идентификации пептидов обычно используются базы данных белков с помощью специализированных поисковых систем, таких как Mascot [2], SEQUEST [3] и X!Tandem [4], которые являются наиболее популярными среди исследователей. Существуют и другие инструменты, например, OMSSA [5] и pFind [6], которые также используются для идентификации пептидов. OMSSA и pFind предлагают уникальные алгоритмы и подходы к обработке и интерпретации масс-спектрометрических данных, что позволяет расширить возможности при работе с биологическими образцами. Эти инструменты анализируют экспериментальные масс-спектрометрические данные, сравнивая их с известными белковыми последовательностями, чтобы определить наиболее вероятные кандидаты в пептиды.

Применение публично доступных онлайн баз данных для расшифровки масс-спектральных данных белков и пептидов значительно ускоряет процесс идентификации состава сложных биологических смесей. Большинство аминокислотных последовательностей, известных на сегодняшний день, собраны в этих базах, обеспечивая исследователям мощный инструмент для анализа и исследования протеомов. Каждая база данных в области биоинформатики имеет свой уникальный подход к организации и хранению данных, обладает разным уровнем детализации и может быть связана с другими схожими ресурсами. Базы данных белков и генетических последовательностей организованы в несколько категорий в зависимости от их функций, структуры и степени взаимосвязи с другими информационными ресурсами. Эти категории помогают выбирать наиболее релевантные источники для исследовательских задач [7]. Первая категория – это базы данных первичного архивирования, например, GenBank, EMBL и Protein Data Bank. Они представляют собой огромные коллекции сырых данных, представленные сообществом учёных и включающие всё от геномных и транскриптомных последовательностей до

структурных данных белков. Вторая категория включает экспертно курируемые базы данных, такие как Swiss-Prot, где каждая запись подвергается тщательной проверке и редактированию специалистами, что обеспечивает высокую точность и надёжность представленной информации. Третий класс состоит из автоматически сгенерированных баз данных, таких как TrEMBL, которые рассчитываются и обновляются с помощью программного обеспечения, применяющего сложные алгоритмы для предварительной классификации и аннотации новых данных, поступающих в базы данных первого типа. Четвёртая группа – это производные базы данных, такие как Structural Classification of Proteins (SCOP) и Protein family database (Pfam), которые предоставляют организованные наборы данных, полученные путём вторичной обработки и анализа информации из первичных и курируемых источников. Наконец, пятая категория объединяет интегрированные базы данных, вроде ENTREZ от NCBI, которые предоставляют унифицированный доступ к широкому спектру информации, собранной из различных баз данных и публикаций.

Поиск по спектральным библиотекам представляет собой передовую методологию в протеомике, которая стремится преодолеть ограничения традиционного секвенирования ДНК. Этот метод использует обширные 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. Это не только повышает скорость анализа, но также улучшает чувствительность и точность определения пептидов, что делает его перспективным подходом в современной протеомике.

Метод секвенирования *de novo* позволяет определять последовательности белков напрямую из масс-спектрометрических данных, исследуя пики, соответствующие фрагментированным пептидным ионам, без использования баз данных белковых последовательностей. Методы секвенирования *de novo*, которые не зависят от баз данных для идентификации пептидов, классифицируются на основе того, как они преобразуют и анализируют данные масс-спектрометрии. Подходы в рамках первой категории опираются на теорию графов и включают стратегии, основанные на определении наилучших путей через спектральные графы для реконструкции последовательности аминокислот. Здесь алгоритмы обрабатывают сложные наборы данных, чтобы выявить пути, которые соответствуют пептидным фрагментам, идентифицируя связи между ионами, образующимися в результате фрагментации молекулы пептида в масс-спектрометре. Во второй категории находятся методы, которые используют статистический анализ и вероятностные модели для предсказания пептидных последовательностей напрямую из масс-спектров. Эти техники включают сложные алгоритмы машинного обучения и искусственного интеллекта, способные обрабатывать большое количество экспериментальных данных для выявления наиболее вероятной аминокислотной последовательности без справочной информации из баз данных. В обоих случаях цель подхода – сопоставить экспериментальные масс-спектры с наиболее вероятной пептидной последовательностью, что позволяет расширить понимание белкового состава и

функциональных изменений. Метод секвенирования *de novo* остаётся ключевым в определении последовательностей пептидов и обнаружении новых белков, а также для выявления мутаций в аминокислотных последовательностях. Однако секвенирование *de novo* является сложным процессом из-за неполноты и несовершенства тандемных масс-спектров. Оптимизация пути в спектре не всегда гарантирует точное определение последовательности пептидов, так как фрагменты пептидов могут быть представлены не в полном виде, а многие пики могут быть результатом помех.

Таким образом, использование баз данных для идентификации пептидов и белков является предпочтительным и широко применяемым методом анализа данных масс-спектрометрии тандемного типа. Этот подход эффективен для распознавания уже изученных белков, чьи последовательности содержатся в базах данных. Однако он ограничен при поиске неизвестных белков или тех, которые претерпели посттрансляционные изменения, что может увеличить время поиска и риск получения неверных результатов. Также одной из сложностей при работе с протеолитическими смесями пептидов является их высокая гомология, которая приводит к генерации множества последовательностей с похожими индексами в результатах поисковых программ. Методология *de novo* секвенирования в масс-спектрометрии сталкивается с конкуренцией множества возможных пептидных последовательностей, претендующих на соответствие одному и тому же спектральному профилю. В таких условиях результаты анализа часто ограничиваются представлением лишь одной последовательности из группы потенциально схожих по структуре пептидов. Этот факт порождает сложности в точности идентификации и подчёркивает необходимость постоянного обновления и калибровки протеомных баз данных для отражения актуальной и корректной информации о белковых последовательностях. В протеомике остро стоит задача обработки масс-спектрометрических данных с учётом вероятных ошибок в существующих аминокислотных последовательностях в базах данных. Решением служит систематическое обновление этих баз, включая верификацию и исправление неточностей на основе новых экспериментальных данных и уточненных компьютерных предсказаний. Секвенирование *de novo* приобретает особую ценность при расшифровке структуры неизвестных пептидов и белков, где точность и полнота спектральных данных становятся критически важными. Эффективность этого подхода возрастает с применением высокоточных масс-спектрометров, способных детально разрешать фрагментные ионы и предоставлять высокую точность идентификации. В этих условиях метод секвенирования *de novo* демонстрирует большой потенциал, позволяя раскрывать новые аспекты белковой функциональности и значительно расширяя понимание протеома.

Достижения в области биоинформатики и других компьютерных алгоритмов помогли открыть огромные объёмы информации о функциях белков с помощью биохимических, биофизических, клеточных биологических и других экспериментальных подходов, интерпретировать результаты протеомных исследований, например, идентифицировать фрагменты белков с помощью масс-

спектрометрии или идентифицировать наличие белков в мультибелковых комплексах, хотя функция белка в этих комплексах часто неизвестна.

Прогнозирование функций белка является ещё одной серьёзной задачей биоинформатики, целью которой является определение функций, выполняемых известным белком. Многие формы данных о белках, такие как белковые последовательности, белковые структуры и сети межбелковых взаимодействий используются для прогнозирования функций. В том числе аминокислотная последовательность белка (так называемая первичная структура) может быть легко определена по последовательности на гене, который её кодирует. Знание этой структуры жизненно важно для понимания функции белка. За последние несколько десятилетий с использованием высокопроизводительных методов было получено большое количество данных о последовательностях белков, что делает их подходящим кандидатом для прогнозирования функций белков с использованием методов глубокого обучения.

Как правило, идентификация функций белка осуществляется посредством ручной или компьютерной аннотации. Первый метод функциональной аннотации, который выполняется экспертами, считается стандартом, так как обеспечивает высокое качество аннотации. Однако из-за его высокой стоимости и трудозатрат он не может обеспечить анализ всего объёма регулярно обновляемых данных. С расширением возможностей технологий секвенирования и ростом числа последовательностей, требующих аннотации, возникла потребность в автоматизированных методах аннотации. Эти методы обеспечивают автоматизацию обработки больших объёмов данных и стремятся повысить точность аннотированных последовательностей.

Были предложены различные базы данных для обеспечения стандартизированной схемы аннотирования функций белков, однако в настоящее время база данных Gene Ontology [8] является наиболее полным ресурсом, поскольку обладает всеми необходимыми свойствами системы функциональной классификации. Консорциум Gene Ontology разработал базу данных, предназначенную для каталогизации и классификации всех функциональных атрибутов, которыми обладают геномные продукты – от отдельных белков до комплексных молекулярных ансамблей. Эта база данных представляет собой контролируемый словарь, включающий в себя описания и определения, связанные с функциональной ролью геномных продуктов в живой клетке. Для обеспечения точности и структурированности информации консорциумом было предложено деление на три категории: молекулярная функция описывает действия на уровне молекул, биологический процесс объединяет эти действия в последовательности, приводящие к определённому результату или изменению состояния, а клеточный компонент связан с местоположением молекулярной активности внутри клетки или в её окружении.

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, зачастую данные о

белковых последовательностях и дополнительная информация из разных биологических баз данных могут храниться в отличных друг от друга форматах, что создаст дополнительные трудности при подготовке входных наборов данных для моделей машинного обучения. Также на этом этапе 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, что делает решение проблемы серьёзным вызовом.

Исходя из вышеизложенного, можно утверждать, что непрерывно растущие объёмы биологических данных предоставляют большие возможности для применения методов машинного обучения в рамках решения разнообразных задач анализа этих данных.

Таким образом, сегодня в сфере обработки генетических и белковых последовательностей большую значимость представляют алгоритмы следующих направленностей: (i) алгоритмы для анализа данных, фокусирующиеся на первичной обработке информации, что облегчает трансформацию сырых спектрометрических данных в формат, пригодный для более глубокого анализа; (ii) специализированные алгоритмы для интерпретации спектральных данных, целью которых является выявление общих паттернов и неявных закономерностей в масс-спектрометрических кластерах и в разнообразии классификационных наборов данных; (iii) базы данных и связанные с ними алгоритмы, которые занимаются запросами идентификации и верификации пептидных и белковых последовательностей, выступая мостом между экспериментальными данными и знанием о структуре и функции биомолекул и (iv) расширенные алгоритмы функциональной аннотации, способные ассоциировать белковые последовательности с биологическими функциями, путём анализа и сопоставления с существующими данными о структурных и функциональных аспектах белков.

Степень изученности и научной разработанности темы исследования.

В рамках диссертационной работы рассматривались публикации авторов по следующим направлениям:

Идентификация пептидов методом секвенирования de novo: Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. [9], Chi H., Chen H., He K., Wu L., Yang B., Sun R., Liu J., Zeng W., Song C., He S., Dong M. [10], Tran N. H., Zhang X., Xin L., Shan B., Li M. [11].

Идентификация пептидов методом поиска в базах данных уже известных белков: Eng J.K., McCormack A.L., Yates J.R., Perkins D.N., Pappin D.J., Creasy D.M., Cottrell J.S., Craig R., Beavis R.C., Brosch M., Yu L., Hubbard T., Choudhary J. [12], Geer L.Y., Markey S.P., Kowalak J.A., Wagner L., Xu M., Maynard D.M., Yang X., Shi W., Bryant S.H. [13], Zhang J., Xin L., Shan B., Chen W., Xie M., Yuen D., Zhang W., Zhang Z., Lajoie G.A., Ma B. [14], Kim S., Mischerikow N., Bandeira N., Navarro J.D., Wich L., Mohammed S., Heck A.J., Pevzner P.A. [15].

Идентификация пептидов методом поиска в спектральных библиотеках: MacLean B., Tomazela D., Shulman N., Chambers M., Finney G., Frewen B., Kern R., Tabb D., Liebler D., Maccoss M. [16], Aebersold R., Lam H. [17].

Предсказание функций белков на основании сходства последовательностей: Piovesan D., Giollo M., Leonardi E., Ferrari C., Tosatto S.C.[18], Gong Q., Ning W., Tian W. [19].

Предсказание функций белков с помощью машинного обучения: Cozzetto D., Minneci F., Currant H., Jones D.T. [20], Jung J., Yi G., Sukno S.A., Thon M.R. [21], You R., Huang X., Zhu S. [22], Yao S., Xiong Y., Sun F. [23].

Предсказание функций белков с помощью глубокого обучения: Rifaioglu A.S., Dogan T., Martin M.J., Cetin-Atalay R., Atalay V. [24], Zhang F., Song H., Zeng M., Wu F-X., Li Y., Pan Y., Li M. [25], Kulmanov M., Hoehndorf R. [26].

Исследования, посвящённые идентификации пептидов методом секвенирования *de novo*, демонстрируют превосходство этого метода для идентификации ранее неизвестных либо модифицированных пептидов. Такие инструменты работают без предварительно аннотированных баз данных белков, что особенно ценно для организмов с неполными геномами. Другим их преимуществом является возможность идентифицировать нестандартные посттрансляционные модификации. Однако эти методы предъявляют очень высокие требования к вычислительным ресурсам, так как требуются интенсивные вычисления, а результаты работы могут быть трудными для интерпретации и потребовать дополнительного подтверждения. Сравнительный анализ показал, что метод секвенирования *de novo* может быть менее точен по сравнению с поиском по базе данных, особенно при низком качестве спектров. Этот метод находит применение в случаях, когда другие методы идентификации неприменимы или недостаточно эффективны, но требует тщательного подхода к анализу данных и проверке результатов.

SEQUEST, один из первых инструментов поиска в базах данных, представленный Eng J.K. и др., является золотым стандартом в современных исследованиях. Он способен коррелировать тандемные масс-спектры с последовательностями аминокислот, хорошо подходит для идентификации пептидов в больших базах данных белков, а также использует кросс-корреляцию для сопоставления теоретических и наблюдаемых спектров. Тем не менее, как и большинство инструментов аналогичного толка, SEQUEST имеет недостатки: он может быть относительно медленным при поиске в базах данных с большим количеством кандидатных пептидов, нуждается в чётких параметрах для посттрансляционных модификаций, которые могут усложнить поиск, а точность идентификации может быть зависимой от качества и разрешения используемых спектров.

Другой инструмент поиска в базах данных, Mascot, впервые был представлен в 1999 году исследователями Perkins D.N. и Pappin D.J. Основным преимуществом Mascot является его способность интегрировать все проверенные методы поиска, включая отпечатки массы пептидов, поиск по последовательности и поиск ионов MS/MS. Методика рассчитывает вероятность совпадения экспериментальных данных с последовательностями из базы

данных, что позволяет получить надёжную степень соответствия. Среди недостатков можно выделить необходимость предопределения фиксированных и переменных модификаций, что может увеличить время поиска и понизить оценки совпадений. Стоит отметить важность настройки таксономии для сокращения времени поиска и улучшения релевантности результатов.

Также важно выделить инструмент X!Tandem, разработанный Craig R. и Beavis R.C. в 2004 году. Этот инструмент способен вычислять статистические оценки для всех индивидуальных сопоставлений спектра с последовательностью, не требует дополнительного программного обеспечения для сборки пептидных сопоставлений и статистического анализа и может обрабатывать экспериментальные данные и выводить результаты в удобном XML формате. Однако X!Tandem сталкивается с типичными для этого класса инструментов проблемами: производительность и масштабируемость могут быть неразрешимой задачей при обработке большого количества данных и многие аналогичные инструменты могут обеспечить лучшую скорость обработки за счёт использования более современных алгоритмов распределения данных и обработки.

Относительно новым считается способ идентификации пептидов с помощью поиска по спектральным библиотекам, описанный в работах MacLean B., а также Aebersold R. Анализ исследований показывает, что данный метод имеет некоторые преимущества перед поиском в базах данных: во-первых, поиск по спектральной библиотеке может быть значительно быстрее, чем традиционный поиск по базе данных белков, поскольку сопоставление происходит с уже идентифицированными спектрами, во-вторых, если спектр уже существует в библиотеке, идентификация может быть более точной, так как исключается необходимость интерпретации сырых данных, в-третьих. эффективность метода высока при анализе образцов, содержащих пептиды, которые уже хорошо представлены в библиотеке. Однако, как и случае с поиском по базам данных, новые или уникальные пептиды могут быть не найдены, так как спектральные библиотеки ограничены спектрами, которые уже были идентифицированы и включены в библиотеку. Также точность идентификации напрямую зависит от полноты и качества спектральных библиотек, которые нуждаются в постоянном обновлении и расширении для включения новых спектров, что является ресурсоёмким процессом.

Подходя к исследованиям, направленным на разработку инструментов функционального аннотирования, в первую очередь стоит отметить ранние инструменты, работающие по методу сходства последовательностей, предложенные Piovesan D. и др. и Gong Q. и др., которые показывают неплохие результаты для выявления функций белков, однако существенно проигрывают по точности аннотации методам, основанным на машинном и глубоком обучении. Их основным недостатком является отсутствие возможности выявлять скрытые закономерности в аминокислотных последовательностях, что делают их устаревшими для работы с большими объёмами биологических данных.

Модели, основанные на машинном обучении, были спроектированы для обнаружения неявных корреляций между многообразными атрибутами белков,

включая их аминокислотную последовательность и трёхмерную структуру. Инструмент FFPred3, работающий на SVM, а также PoGO, который использует методы, основанные на опорных векторах и линейной классификации, демонстрируют хорошие результаты аннотирования, но эти инструменты используют дополнительные атрибуты белков, что существенно ограничивает их диапазон действия, так как большая часть неаннотированных белков не содержит никак информации кроме их аминокислотных последовательностей.

На текущий момент последним веянием в задаче аннотирования белков являются методы глубокого обучения. Среди работ по решению проблемы функционального аннотирования можно выделить DEEPred, предложенный Rifaioglu A.S. и др. в 2019 и основанный на DNN, DeepGOA, разработанный Zhang и др. в тот же период, основанный на BiLSTM и использующий информацию о PPI, а также инструмент 2020 года DeepGOPlus авторов Kulmanov M. И Hoehndorf R., использующий CNN в своей основе. Данные инструменты показывают отличные результаты предсказаний и являются очень перспективными для дальнейших исследований. Однако и они не лишены недостатков, таких как использование дополнительных атрибутов белков и невозможность эффективно фиксировать долгосрочную зависимость между одной и той же последовательностью белка.

Таким образом, все представленные работы авторов обладают как преимуществами, так и недостатками, многие из которых не удаётся преодолеть до сих пор. В представленной диссертационной работе предлагаются алгоритмы, направленные на решение существующих проблем как в области идентификации пептидов, так и в области функционального аннотирования. Алгоритмы используют только мотивы последовательности, что позволяет работать практически с любыми наборами данных пептидов и белков, не сталкиваясь с проблемой отсутствующей или неполной дополнительной информации о белок-белковых взаимодействиях, третичной структуре и т. п.

Отличие представленных алгоритмов состоит в их архитектуре, использующей современные подходы в глубоком обучении. В частности, в рамках решения задачи идентификации пептидов представлена модель, использующая изучение функции кросс-модального сходства, которая встраивает спектры и пептиды в общее евклидово подпространство для прямого сравнения. В рамках решения задачи предсказания функций белков предложена модель BiLSTM, скомбинированная с механизмом самовнимания, который напрямую связывает корреляцию между любыми двумя функциями в последовательности

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 Владимира Сергеевича, PhD, Научно-исследовательская платформа сельскохозяйственной биотехнологии, а также данные из общедоступных баз данных белковых и ДНК-последовательностей, в том числе NIST, Pfam и UniProt.

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

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.

Материалы исследования:

1. Данные масс-спектрометрии, полученные в рамках проекта «ПЦР-тест для детекции и дифференциальной диагностики возбудителей описторхоза и меторхоза» под руководством Киян Владимира Сергеевича, PhD, Научно-исследовательская платформа сельскохозяйственной биотехнологии (НИПСБ).
2. Данные из баз данных белковых и ДНК-последовательностей NIST, PRIDE, Pfam, GenBank и UniProtKB/Swiss-Prot.

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.

Научная новизна диссертационного исследования:

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

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.

Теоретическая значимость результатов диссертационного исследования:

1. Разработанные алгоритмы не имеют ограничений на длину аминокислотной последовательности и, следовательно, могут использоваться для аннотации функций белка в масштабе генома;

2. Работают быстро и могут аннотировать несколько тысяч белков за несколько минут даже на одном процессоре;

3. Модели не ограничены несбалансированной или отсутствующей информацией о межбелковых взаимодействиях;

4. Алгоритмы можно применять, используя только мотивы последовательности.

Практическая значимость результатов диссертационного исследования:

1. Разработанные алгоритмы могут быть внедрены программные модули лабораторий для идентификации белковых последовательностей и предсказания их функций с высокой надёжностью.

2. Разработанные алгоритмы могут быть использованы биологами как альтернативный существующим приложениям либо дополнительный инструментарий при работе с биологическими данными.

3. 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» 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», КАТУ им. С. Сейфуллина, Нур-Султан, 2020 год [27];

- Международная научно-практическая конференция «Интеграция науки, образования и производства основа реализации Плана Нации», КарГТУ, 2020 год [28];

- Международная научно-теоретическая конференция «Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация»», Нур-Султан, 2021 год [29];

- Международная научная конференция «XXII Сатпаевские чтения», Satbayev University, Алматы, 2022 год [30];

- Международная научно-теоретическая конференция «Сейфуллинские чтения – 18: «Молодёжь и наука – взгляд в будущее»», КАТУ им. С. Сейфуллина, Астана, 12 апреля 2022 год [31];

- Международная научная конференция «Математическая логика и компьютерные науки», ЕНУ им. Л. Н. Гумилёва, Астана, 7-8 октября 2022 год [32].

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 (Приложение А).

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 с использованием инструментов для выделения ключевых моментов, таких как иллюстрации, схемы и таблицы, сопровождается 23 рисунками и 7 таблицами.

Во введении подчёркивается важность изучаемой темы, уровень изученности и научной проработанности, раскрыта актуальность разработанных алгоритмов для обработки данных протеомики, сформулирована цель исследования, поставлены задачи для достижения цели, определены предмет и объект исследования, раскрыты научная новизна, теоретическая и практическая значимость диссертационного 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 ранние и современные методы для решения задачи идентификации белков и пептидов. Помимо этого, осуществлён анализ проблематики аннотации белков, полученных экспериментальным путём, и критический обзор алгоритмов, предложенных предшественниками для

функционального прогнозирования. Выявлены ключевые недостатки и намечены направления для усовершенствования этих алгоритмов для повышения их точности и надёжности.

Во втором разделе предложено решение для идентификации пептидов и белков на основе общедоступной сети глубокого подобия SpeCollate, которая была адаптирована для создания вложений единого размера спектров и пептидов в единое евклидово пространство, что позволяет обеспечить их сравнение путём оптимизации и развития сети. Для достижения этой цели был выбран определённый набор параметров модели, включающий использование двунаправленной LSTM. Также охарактеризован процесс обучения сети, который заключается в обучении 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 математическая 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 по общепринятым параметрам, используемым для анализа точности и надёжности моделей машинного обучения. Также приведены результаты ручного аннотирования функций белка, подтверждающие высокую точность разработанного алгоритма.

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

В приложениях представлены авторские свидетельства и акты внедрения.

Автор выражает глубокую благодарность своему научному консультанту доктору философии, ассоциированному профессору кафедры «Информационные системы» Исмаиловой Айсулу Абжаппаровне за неоценимую помощь в ходе диссертационного исследования. Автор признателен своему зарубежному консультанту кандидату физико-математических наук, старшему научному сотруднику ИСИ СО РАН (Российская Федерация, г. Новосибирск) Штокало Дмитрию Николаевичу за консультации в ходе исследований.

Также автор выражает благодарность за предоставленные практические материалы, экспериментальные данные и консультации в вопросах протеомики и генетики доктору философии Киян Владимиру Сергеевичу.

1 ИССЛЕДОВАНИЕ ПРОБЛЕМЫ АНАЛИЗА ДАННЫХ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

1.1 Анализ текущего состояния международных баз данных белковых и генетических последовательностей

Биологические базы данных (ББД) представляют собой цифровые архивы, собирающие, структурирующие и интегрирующие разноплановые данные о биологических объектах, таких как генетические последовательности, белковые структуры, клеточные и молекулярные функции, а также метаболические и сигнальные пути. Они служат краеугольным камнем современной биологии и биомедицины, обеспечивая поддержку всего спектра научных исследований от фундаментальной науки до клинических применений.

Биологические базы данных являются незаменимым инструментом, позволяющим хранить и анализировать огромные объёмы генетической информации. ББД облегчают совместную работу учёных со всего мира и ускоряют открытие новых знаний, что, в свою очередь, способствует развитию медицины, фармакологии, сельского хозяйства и других важных областей. На сегодняшний день насчитывается несколько тысяч международных биологических баз данных. Множество новых баз данных создаются в разных областях биологии и медицины для удовлетворения специфических исследовательских нужд. Можно выделить следующие основные категории биологических баз данных:

- базы данных генетических последовательностей,
- базы данных белков,
- базы данных метаболических и биохимических путей,
- базы данных экспрессии генов и транскриптомики,
- базы данных фенотипов и заболеваний,
- базы данных таксономии и биоразнообразия,
- интегративные базы данных, содержащие агрегированные данные из разных источников.

Существуют также специализированные базы данных, предназначенные для узких областей исследований, таких как нейронаука, экология, сельское хозяйство и другие.

В диссертационном исследовании рассмотрено текущее состояние наиболее крупных и регулярно обновляемых международных баз данных белковых и ДНК последовательностей. Были изучены и проанализированы следующие репозитории: GenBank, UniProt, Protein Data Bank (PDB), European Nucleotide Archive (ENA) и Pfam. Во внимание принимались такие показатели как количество данных, частота обновления, процент прироста новых данных в разрезе разных периодов, участие курирующих экспертов, форматы представления данных и их доступность, взаимодействие с другими базами данных, наличие инструментов для доступа к данным и их анализа.

База данных GenBank [33] является одной из наиболее крупных и значимых баз данных ДНК последовательностей. Созданная в 1982 году, она в настоящее время содержит более 400 миллионов записей, включающих

информацию о геномах, транскриптомах, а также других молекулярных последовательностях. GenBank является частью Международного сотрудничества по базам данных нуклеотидных последовательностей, которое включает в себя Банк данных ДНК Японии (DDBJ), Европейский архив нуклеотидов (ENA) и GenBank в Национальным Центре Биотехнологической Информации (NCBI). Эти три организации ежедневно обмениваются данными.

Далее представлен обзор ключевых характеристик GenBank:

1. GenBank регулярно обновляется, причём новые данные добавляются каждые два месяца. Эти данные поступают из научных статей, напрямую от исследователей и больших проектов, таких как проект Генома Человека.
2. GenBank предоставляет обширную библиотеку генетических данных, доступную пользователям по всему миру через удобные онлайн-сервисы и инструменты, такие как Entrez, система поиска NCBI, и BLAST, инструмент для локального сравнения последовательностей. Эти мощные инструменты аналитики позволяют проводить глубокий поиск среди базы данных последовательностей, обнаруживать сходства, анализировать эволюционные связи и расшифровывать функциональные области внутри белков и генов.
3. GenBank тесно взаимодействует с другими международными базами данных последовательностей, такими как EMBL в Европе и DDBJ в Японии. Они обмениваются данными, чтобы каждая последовательность была представлена во всех трёх ресурсах.
4. Кроме последовательностей, GenBank также содержит связанную информацию, такую как аннотации, публикации, локализацию генов, белковые продукты и комментарии.
5. GenBank служит основным источником данных для многих биологических исследований и приложений, включая таксономию, филогенетику, молекулярную биологию и эволюционную биологию.
6. GenBank предлагает данные в нескольких форматах, которые поддерживаются различными биоинформатическими инструментами, включая формат GenBank, FASTA и другие.

Современное состояние GenBank отражает его постоянное развитие и адаптацию к потребностям научного сообщества. GenBank регулярно обновляется, и каждый новый релиз включает данные с последующими обновлениями, доступными через инкрементные файлы обновлений. Это обеспечивает текущее обновление данных в базе и предоставляет пользователям возможность загрузки самой актуальной информации.

GenBank имеет ряд преимуществ и особенностей, которые делают её одной из наиболее востребованных баз данных в области геномики и биоинформатики. Одним из основных преимуществ GenBank является высокая точность и качество информации, которая хранится в базе данных. Для того чтобы гарантировать высокое качество данных, GenBank использует строгие процедуры контроля качества, которые позволяют идентифицировать и устранять ошибки в ДНК последовательностях.

Кроме того, база данных GenBank содержит огромное количество информации о геномах различных организмов, включая людей, животных, растения и микроорганизмы. Это делает GenBank важным инструментом для исследований в области геномики, эволюции и биологии развития. База данных также содержит множество полезных инструментов и программ для работы с данными, включая программы для анализа последовательностей, поиска генов и выравнивания ДНК последовательностей.

GenBank, как одна из ведущих генетических баз данных, сталкивается с рядом технических и организационных проблем, связанных с быстрым ростом данных, в первую очередь это может создавать сложности при хранении, поиске и доступе к данным. Ещё одной важной проблемой является неоднородность данных. Данные, поступающие в GenBank, могут значительно отличаться по своему качеству и формату. Это связано с различиями в методиках секвенирования, аннотации и стандартах представления информации от разных источников. Неоднородность может затруднять автоматизированный поиск и сравнительный анализ данных.

Также по мере увеличения объёма данных усилия по их курированию становятся всё более трудоёмкими, что приводит к низкому качеству аннотации. Несмотря на использование автоматических систем для первоначальной аннотации последовательностей, важность ручного курирования и проверки остаётся высокой, так как точные аннотации крайне важны для понимания биологической функции и контекста последовательностей.

Отдельная проблема – присутствие загрязнённых последовательностей, таких как случайно включённые некорректные или бактериальные последовательности в геномах более крупных организмов. Это может привести к ошибочным результатам при анализе данных. Например, было обнаружено, что в геномах человека и некоторых модельных организмов присутствуют участки, соответствующие бактериальным последовательностям, что указывает на потенциальное загрязнение данных. Это требует дополнительных усилий по курированию и очистке базы данных.

В целом, база данных GenBank является важным инструментом для исследований в области геномики и биоинформатики. Благодаря высокому качеству и доступности данных, а также широкому спектру инструментов и приложений, GenBank позволяет исследователям по всему миру проводить качественные исследования в области геномики, эволюции и биологии развития. Однако, необходимо учитывать проблемы, связанные с объёмом данных и качеством аннотаций, и постоянно совершенствовать инструменты и методы работы с данными в GenBank.

UniProt, или Universal Protein Resource [34], является крупнейшим хранилищем данных о белках. Он предоставляет централизованную ресурсную базу данных, содержащую обширную коллекцию информации о белках, включая их функции, классификацию, доменную структуру и последовательности.

UniProt была создана в 2003 году путём объединения баз данных Swiss-Prot, TrEMBL и PIR. Сегодня UniProt является центральным ресурсом для информации о белках и содержит данные о более чем 200 миллионах белков,

собранных из более чем 20 тысяч организмов. Кроме того, база данных UniProt постоянно обновляется и содержит информацию о последних научных открытиях в области белков.

Ключевые характеристики базы данных UniProt:

1. UniProt включает в себя несколько баз данных, среди которых UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef) для уменьшения избыточности и ускорения последовательностного анализа, и UniProt Archive (UniParc), которая служит центральным хранилищем белковых последовательностей.
2. UniProtKB – это основная база данных в UniProt, она делится на две секции: UniProtKB/Swiss-Prot, которая содержит ручную аннотацию и проверку курирующими экспертами, и UniProtKB/TrEMBL, содержащую автоматически аннотированные записи, ожидающие рецензирования.
3. Данные в UniProt регулярно обновляются, что обеспечивает доступ к наиболее актуальной информации о белковых последовательностях и функциях.
4. UniProt предоставляет свободный и открытый доступ к своим данным через веб-сайт и программное обеспечение для поиска и извлечения данных.
5. UniProt работает в тесной связи с множеством других биологических баз данных и интегрирует информацию из различных источников для предоставления всесторонней информации о белках.

К списку проблем, с которыми сталкивается UniProt, можно отнести следующие пункты: поддержание и обновление такого обширного ресурса требует значительных финансовых вложений; качество аннотаций – несмотря на автоматизированные подходы к аннотации, качество и точность данных могут варьироваться, и ручное курирование остаётся важным для обеспечения надёжности информации. Также важно обозначить общую проблему всех биологических баз данных – быстро растущий объём данных: обработка и управление огромным количеством данных, особенно в свете растущего числа геномных проектов, является сложной задачей. Отсюда вытекает проблема интеграции с другими ББД, а также стандартизация данных для обеспечения совместимости и возможности сравнения между разными видами и экспериментами.

Одной из главных задач базы данных UniProt является обеспечение точности и актуальности данных. Для этого база ежедневно обновляется с помощью автоматических и ручных методов анализа. Эти методы включают в себя использование данных из других репозиториев, ручную проверку аннотаций и экспертную оценку. База данных UniProt также предоставляет различные инструменты для анализа данных. Эти инструменты могут быть использованы для поиска белков по различным параметрам, для анализа последовательностей и структур белков, а также для анализа генетических вариантов, связанных с заболеваниями.

UniProt также является частью множества других баз данных и ресурсов, связанных с белками. Например, UniProt сотрудничает с базами данных Protein

Data Bank и InterPro, предоставляя возможность быстрого доступа к дополнительной информации о структуре белков и их функциях.

В целом, UniProt остаётся критически важным ресурсом для глобального научного сообщества, предоставляя фундаментальные данные для множества биологических исследований и приложений. Продолжающиеся усилия по улучшению качества, доступности и функциональности UniProt важны для поддержания его статуса как ведущего информационного ресурса в области биоинформатики.

В рамках данного исследования важно выделить и более подробно описать *UniProtKB* (UniProt Knowledgebase), которая является центральной базой данных проекта UniProt, предназначенной для хранения информации о последовательностях и функциях белков. Она является одним из наиболее важных инструментов для биологов и биоинформатиков по всему миру. Включает в себя информацию из различных источников, в том числе геномные проекты, эксперименты по масс-спектрометрии белков и научные статьи.

UniProtKB делится на два основных раздела:

1. UniProtKB/Swiss-Prot – раздел с ручной аннотацией. Записи в Swiss-Prot содержат подробную, проверенную и точно аннотированную информацию о белках. Каждая запись проходит тщательный процесс рецензирования экспертами, что обеспечивает высокую точность и надёжность представленной информации. Эта часть базы данных является менее обширной по сравнению с TrEMBL, но предлагает более высокое качество аннотаций.
2. UniProtKB/TrEMBL – этот раздел представляет собой коллекцию автоматически аннотированных записей, которые ожидают ручной обработки и перевода в раздел Swiss-Prot. Аннотации создаются на основе информации из различных источников и предсказательных программ, из-за чего могут содержать менее точные данные.

Protein Data Bank (PDB) [35] является базой данных, содержащей информацию о трёхмерных структурных данных белков, нуклеиновых кислот и комплексов, которые обычно определяются с помощью таких методов структурной биологии, как рентгеновская кристаллография, ядерно-магнитный резонанс и криоэлектронная микроскопия (cryo-EM). PDB управляется совместно несколькими организациями в рамках Всемирного консорциума по банку данных белков. На сегодняшний день в PDB содержится более 180 000 структурных записей [36]. Большинство записей относится к белкам, но также в PDB содержится информация о нуклеиновых кислотах и других макромолекулах.

База данных Protein Data Bank была создана в 1971 году как первая база данных белковых структур. Одним из главных достижений PDB является создание стандарта формата для хранения информации о структурах белков. Данный формат называется PDB-форматом и является стандартом для хранения структурных данных во всём мире. Благодаря этому формату исследователи и учёные могут легко обмениваться информацией о структурах белков и использовать её в своих исследованиях.

PDB имеет огромную значимость для многих областей науки, включая биологию, медицину, фармакологию и биоинформатику. Благодаря PDB учёные могут изучать структуру белков и понимать их функции, что позволяет разрабатывать новые лекарственные препараты и методы лечения различных заболеваний. Кроме того, PDB является важным инструментом для разработки новых методов биоинформатики и машинного обучения.

Основные характеристики Protein Data Bank:

1. Открытый доступ. Данные PDB доступны всем пользователям бесплатно, и их можно свободно скачивать через веб-порталы или программные интерфейсы.
2. Структуры, содержащиеся в PDB, стандартизированы в формате файлов PDB, который содержит подробную информацию о положении каждого атома в белке или нуклеиновой кислоте, а также о химических связях между ними.
3. PDB акцентирует внимание на качестве и точности структурных данных, включая проверку структур на соответствие экспериментальным данным.
4. Пользователям предоставляются различные инструменты и программное обеспечение для визуализации трёхмерных структур, что позволяет лучше понять молекулярное строение и функции белков и других макромолекул.
5. PDB служит важным образовательным ресурсом для студентов и исследователей, заинтересованных в структурной биологии и молекулярном моделировании.

Однако, несмотря на все преимущества PDB, база данных также имеет свои ограничения и недостатки:

- проблема управления данными: большие объёмы структурных данных требуют эффективного управления и хранения, а также обеспечения быстрого доступа к ним,
- по мере того, как совершенствуются методы структурной биологии, появляются новые данные, которые могут заменять или дополнять существующие структуры, требуя постоянного обновления базы данных,
- точная аннотация структур, включая функциональные сайты, взаимодействия и динамику молекул, требует усилий по курированию и обновлению данных.

PDB представляет собой ключевую информационную платформу, которая вносит значимый вклад в биомедицинские науки, предоставляя всесторонние данные о трёхмерной структуре макромолекул. Сведения, хранящиеся в PDB, включают в себя детализированные модели белков, ДНК и РНК, и эти модели имеют первостепенное значение не только для разгадывания биологических механизмов на молекулярном уровне, но и для проектирования и разработки новых терапевтических препаратов. Эффективность и надёжность PDB подкрепляется строгими процедурами качественной аннотации и постоянным обновлением данных, совершаемыми совместно в рамках международного сотрудничества. Тем не менее, PDB сталкивается с вызовами, связанными с

управлением растущим объёмом сложных данных и поддержанием их качества. Непрерывные усилия по стандартизации и верификации данных необходимы для того, чтобы PDB оставалась актуальной, точной и полезной для исследователей по всему миру.

База данных ENA (European Nucleotide Archive) [37] является одной из ведущих баз данных, содержащих информацию о нуклеотидных последовательностях ДНК, РНК и белков. База данных ENA является частью системы EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) и содержит множество данных, собранных разными исследовательскими группами по всему миру. На момент написания исследования в ENA было более 1.5 миллиарда записей.

Основным функционалом базы данных ENA является хранение, анализ и обработка нуклеотидных последовательностей, связанных с генетическими исследованиями. Это включает в себя геномы, транскриптомы, метагеномные данные, метаболомные данные и другие типы последовательностей.

Данные в ENA представлены в плоском файловом формате, известном как EMBL-Bank format, который использует уникальный синтаксис для представления записей. Для удобства пользователей данные доступны через различные форматы, включая XML, HTML, FASTA и FASTQ, и могут быть извлечены вручную или программно через браузер ENA.

ENA также оперирует экземпляром Sequence Read Archive (SRA), который является архивом последовательностей чтения секвенсоров следующего поколения и анализов, предназначенных для публичного раскрытия. SRA быстро растёт и в настоящее время принимает последовательности чтения, сгенерированные с помощью платформ секвенирования нового поколения, а также некоторые соответствующие анализы и выравнивания.

Одной из особенностей базы данных ENA является её открытость и доступность. Любой исследователь может загрузить свои данные в базу данных ENA, что позволяет создать большую и разнообразную коллекцию данных для научных исследований.

Кроме того, база данных ENA предоставляет широкий функционал для анализа данных. Например, в базе данных доступны инструменты для поиска и сравнения последовательностей, а также для анализа их функций и структуры.

База данных ENA также является частью международных инициатив, таких как International Nucleotide Sequence Database Collaboration (INSDC), которая включает в себя сотрудничество с другими базами данных, такими как GenBank и DDBJ.

Одной из ключевых проблем для ENA является управление огромными объёмами данных, что представляет значительную проблему хранения. Для управления этим ростом ENA применяет стратегии сжатия данных и, по мере необходимости, может отказываться от данных менее ценных платформ секвенирования.

ENA играет важную роль в поддержании глобальной научной инфраструктуры, позволяя учёным обмениваться и использовать последовательности нуклеотидов в их исследованиях. Это позволяет проводить

сравнительные исследования в различных областях, от молекулярной биологии до эволюционной генетики, и поддерживать глобальные инициативы.

База данных Pfam (Protein Families Database) [38] является одним из наиболее известных и широко используемых инструментов для анализа белковых последовательностей и классификации белковых семейств. Она содержит информацию о более чем 18 000 семейств белков и более чем 20 миллионов последовательностей.

Основной целью базы данных Pfam является описание семейств белков, включая определение их структурных и функциональных характеристик. Данные из этой базы данных используются во многих областях, включая биоинформатику, молекулярную биологию, фармакологию и другие.

Состояние базы данных Pfam постоянно обновляется и развивается. На момент написания диссертационного исследования база данных содержит информацию о более чем 18 000 семействах белков [39], которые были выделены на основе их схожести в последовательностях и структурных особенностях. Каждое семейство имеет свой уникальный идентификатор, который позволяет быстро найти все связанные с ним данные в базе данных.

Основные характеристики Pfam:

1. Каждое белковое семейство в Pfam имеет курированное выравнивание, которое содержит репрезентативный набор последовательностей для данного семейства.
2. Для каждого семейства создаётся скрытая марковская модель, что позволяет идентифицировать эти домены в новых белковых последовательностях.
3. Pfam предлагает различные веб-инструменты для анализа и исследования белковых семейств, включая поиск по базе данных и инструменты для визуализации структур.
4. В Pfam также есть материалы и руководства, которые могут быть полезны для обучения и развития навыков в области биоинформатики и структурной биологии.

Кроме того, база данных Pfam также содержит информацию о функциях белков, включая информацию о доменах, мотивах и других структурных элементах, которые могут влиять на их функционирование. Это позволяет ученым быстро определить, какие белки могут выполнять схожие функции в разных организмах.

Одним из наиболее значимых преимуществ базы данных Pfam является её интеграция с другими базами данных, такими как UniProt, NCBI и другие. Это позволяет пользователям быстро и легко получать доступ к большому количеству данных о белках и использовать их для дальнейших исследований. Благодаря постоянному обновлению и развитию, она остаётся одним из наиболее важных ресурсов для биоинформатики и молекулярной биологии.

Результаты исследования показывают, что рост и эволюция международных баз данных белковых и генетических последовательностей являются ключевыми для поддержания темпов современных биологических

исследований. Важным направлением развития является улучшение качества и точности данных, а также улучшение методов анализа.

Интеграция данных из разных источников представляет собой другую значительную проблему, так как разные репозитории часто используют несовместимые форматы и стандарты. Разработка универсальных инструментов интеграции, которые могут сглаживать эти различия и обеспечивать эффективный обмен данными, является ключевым направлением для усиления исследований и использования данных в биоинформатике. В результате необходимо использовать алгоритмы, которые могут эффективно интегрировать эти данные вместе и предоставлять исследователям доступ к всем этим данным в едином формате. С ростом объёма данных, который представляет собой как возможность, так и вызов, становится очевидной необходимость в разработке новых и более мощных вычислительных инструментов для их обработки. Эти инструменты должны быть способны не только справляться с текущими объёмами данных, но и быть масштабируемыми для будущих расширений.

Другим вызовом является необходимость обрабатывать данные, которые были получены с использованием новых технологий. Например, в последние годы стали широко использовать технологии секвенирования следующего поколения, которые позволяют быстро получать большое количество данных о генетических последовательностях. Однако, эти данные могут быть сложными для обработки из-за их большого объёма и высокой степени зашумлённости. Кроме того, существует необходимость разработки алгоритмов, которые могут работать с неоднородными данными, такими как данные из различных экспериментов и с разными параметрами. Эти данные могут иметь разную точность и качество, что может также усложнить их обработку.

Наконец, одной из ключевых задач является выявление скрытой информации, которая может иметь значение для понимания биологических процессов и заболеваний. Например, белки и ДНК последовательности могут быть связаны с определёнными заболеваниями или фенотипами, но эта информация может быть не очевидна из первоначальных данных. Для обнаружения этой информации необходимы новые алгоритмы, которые могут интегрировать различные источники данных и выявлять связи между ними. Для того чтобы открыть эти скрытые связи, требуются сложные аналитические методы, способные синтезировать и анализировать данные из разных источников. Это могут быть алгоритмы машинного обучения, статистический анализ, сетевой анализ или комбинация этих и других методов. Эти алгоритмы должны быть способны обрабатывать большие объёмы данных и интегрировать информацию на множественных уровнях, включая генетическую, метаболическую и фенотипическую информацию. Новые подходы к интеграции данных могут включать разработку мультимодальных платформ, где данные о последовательностях сочетаются с клиническими данными, данными экспрессии генов и белковых взаимодействиях. Это позволит исследователям не только найти скрытые связи между последовательностями и заболеваниями, но и понять механизмы, которые лежат в основе этих связей.

Таким образом, в свете растущей сложности и размера данных текущее состояние баз данных белковых и ДНК последовательностей приводит к необходимости развития методов анализа, включая алгоритмы машинного обучения и статистический анализ. Эти инструменты позволяют исследователям распознавать сложные закономерности и связи, которые могут быть неочевидны при традиционных подходах.

1.2 Исследование проблемы достоверной идентификации пептидов и оценки её точности

Статистически точная идентификация белка является фундаментальным краеугольным камнем протеомики и лежит в основе понимания и применения этой технологии во всех областях медицины и биологии [40].

Пептиды, которые были идентифицированы с помощью различных систем поиска последовательностей, часто неверны и требуют тщательной статистической проверки совпадений пептидного спектра (Peptide-Spectrum Matches – PSM) перед дальнейшим анализом. Таким образом, для определения правильных идентификаций с высокой степенью достоверности существуют различные схемы оценки, некоторые из которых уже включены в поисковые инструменты (Sequest, Mascot, X!Tandem). Эти баллы отображают меру сходства либо между пептидной последовательностью и спектрами, либо между наблюдаемыми и библиотечными спектрами. Баллы обычно преобразуются в более подробную статистическую оценку, и для каждого полученного балла соответствия рассчитываются вероятности или *e*-значения для проведения оценки частоты ложных открытий (False Discovery Rates – FDR) [41]. Двумя широкими категориями оценки FDR для спектров MC/MC являются поиск цели-приманки [42] и эмпирический байесовский подход.

Создание списка идентифицированных белков на основе пептидных последовательностей с определёнными идентификационными баллами сталкивается с рядом проблем, которые могут включать в себя сложности в интерпретации спектров, недостаточную представленность некоторых пептидов в базах данных и потенциальные ошибки из-за посттрансляционных модификаций. Обозначим несколько ключевых моментов:

1. Идентификация белков на основе пептидных последовательностей сталкивается с проблемами, особенно когда доступное количество идентификаций пептидов ограничено и часто оказывается ненадёжным [43]. Такое положение дел возникает из-за предпочтения, отдаваемого пептидно-спектральным совпадениям с высокими баллами, из которых лишь немногие могут быть признаны достоверными. Это затрудняет подтверждение идентификации белков, особенно если для белка подтверждается только один пептид.
2. Для надёжной идентификации белка обычно требуется несколько пептидов, потому что идентификация на основе одиночного пептида может быть недостоверной [44, 45]. Это означает, что в результатах поиска по базам данных часто оказывается лишь ограниченное количество пептидных совпадений для каждого белка, и только кандидаты с высокими идентификационными баллами

считаются достоверными. Пептиды одного белка также могут иметь различную вероятность обнаружения в масс-спектрометрических экспериментах, что зависит от их детектируемости [46], или возможности их обнаружения в данных условиях эксперимента.

3. Многие пептидные последовательности, встречающиеся в типичном рабочем процессе протеомики, могут быть сопоставлены более чем с одним белком в базе данных. Их называют вырожденными или общими пептидами [47].

4. Оценка частоты ложного обнаружения идентифицированных пептидов и белков также является весьма нетривиальной задачей. Некоторые подходы к оценке FDR на уровне пептидов включают создание баз данных ложных пептидов или неконтролируемую оценку условных распределений классов (распределения оценок PSM при правильной и ложной идентификации, соответственно). Однако большое количество PSM с низкой оценкой может создавать трудности в определении достоверности идентификации как пептидов, так и белков.

На сегодняшний день было разработано несколько методов и стратегий для интерпретации пептидов в белки, например сопоставление больших комбинированных непрерывных спектров МС/МС непосредственно с последовательностью белка в базе данных и группирование последовательностей пептидов в белки с помощью статистических методов (путём присвоения вероятностей и баллов). После вывода данных о белке снова выполняется статистическая проверка, и FDR оцениваются на уровне белка. Хотя методы оценки FDR на уровне пептидов хорошо описаны, вычисление FDR на уровне белков остаётся открытой проблемой [48].

Таким образом, на данный момент нет универсального стандарта для оценки эффективности методов как поиска пептидов в базах данных белков, так и секвенирования *de novo*, и точность таких методов может сильно варьироваться. Существует множество подходов, включая поиск по базе данных и *de novo* алгоритмы, каждый из которых имеет свои слабые стороны, включая вопросы точности идентификации и общую статистическую надёжность. Это подчёркивает необходимость дальнейших исследований и разработки более надёжных методов идентификации и оценки её качества.

1.3 Анализ существующих алгоритмов и решений для идентификации белковых последовательностей

Обычно выделяют три метода идентификации пептидов, включая секвенирование *de novo*, поиск в спектральной библиотеке и поиск в базе данных белков. Секвенирование *de novo* извлекает интервалы пиков из спектров МС/МС и использует их для построения рядов ионов *b* или *y*, выводя, таким образом, целые пептидные последовательности. Однако из-за неполных рядов ионов *b* и *y* в большинстве спектров этот метод редко может дать полную последовательность пептидов и в основном используется для интерпретации спектров, которые не удалось найти в базе данных, или спектров неизвестных видов.

Поиск по библиотеке спектров – это недавно разработанный метод, аналогичный поиску спектров фрагментации малых молекул. Большое количество идентифицированных экспериментальных спектров хранится вместе с их последовательностями в виде библиотеки, и новый экспериментальный спектр нужно только сравнить с этими сохраненными спектрами, чтобы определить наилучшее совпадение. Этот метод предлагает превосходную скорость и чувствительность, однако основным ограничением является то, что его можно применять только к известным спектрам, но он не позволяет делать новые открытия. Самая популярная спектральная библиотека доступна в Национальном институте стандартов и технологий [49].

Поиск в базе данных белков является наиболее распространённым способом идентификации пептидов. Идея состоит в том, чтобы генерировать теоретические спектры из доступных белковых последовательностей и сравнивать их с экспериментальными спектрами. Легко представить, что теоретические спектры не так точны, как спектры в библиотеке, но белковые базы данных всегда охватывают достаточно большое количество белков, либо обнаруженных в предыдущих экспериментах, либо непосредственно аннотированных из генома.

Наиболее часто используемые алгоритмы идентификации белков предназначены для идентификации последовательностей по спектрам фрагментации, полученным диссоциацией, индуцированной столкновением (collision induced dissociation – CID), при которой ионы-предшественники пептидов сталкиваются с молекулами инертного газа и диссоциируют. CID обычно приводит к фрагментации пептидного остова по амидным связям с образованием преимущественно *N*-концевых ионов *b* и *C*-концевых *y*. Другие типы ионов, включая потери нейтральной воды и аммиака, а также расщепление боковых цепей, также возможны, но менее распространены. Поскольку массы ионов-продуктов предсказуемы, последовательность исходного пептида может быть реконструирована из спектра МС/МС путём сопоставления экспериментальных масс ионов фрагментов с теоретическими. Долгое время *m/z* были основной информацией, используемой популярными алгоритмами, включая Sequest, X!Tandem и Mascot, для сопоставления пептидных последовательностей со спектрами фрагментации. Процесс состоит из поиска в базе данных белков или базе данных транслированных нуклеотидов по *m/z* в пределах определённого допуска относительно предшественника *m/z* для возможных пептидов-кандидатов. После идентификации кандидатов каждый экспериментальный спектр сравнивается со многими сконструированными теоретическими спектрами или списками пиков, которые соответствуют последовательностям пептидов-кандидатов, и каждой последовательности-кандидату присваивается оценка на основе сходства между теоретическим и экспериментальным спектрами или вероятности, что их совпадение не является случайным. Процесс идентификации белков с помощью баз данных продемонстрирован на рисунке 1.

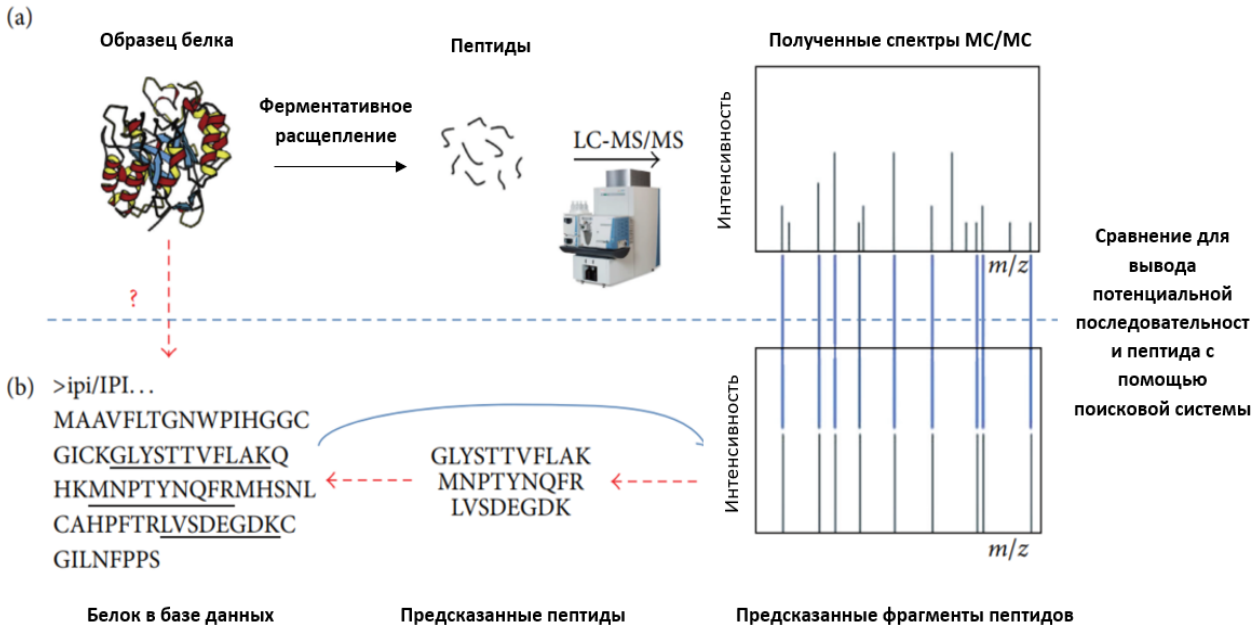


Рисунок 1 – Процесс идентификации пептидов

Далее будут описаны несколько наиболее часто используемых алгоритмов поиска в базе данных, включая SEQUEST, Mascot, X!Tandem, OMSSA, Phenyx, PEAKS DB и MS-GFDB.

SEQUEST – это алгоритм, который выполняет сопоставление заданного необработанного спектра МС/МС с потенциальными последовательностями путём применения методик оценки и сортировки, которые базируются на спектральном сходстве. Это достигается за счёт сравнения теоретически предсказанных спектров с фактически наблюдаемыми в экспериментальных данных. Стратегия анализа начинается с компьютерной обработки данных тандемной масс-спектрометрии (Шаг 1). Затем последовательности аминокислот идентифицируют в базе данных белков для сравнения с обработанными данными тандемной масс-спектрометрии путём сопоставления молекулярной массы пептида с линейной последовательностью (Шаг 2). Предсказанные ионы фрагментов последовательностей, полученных из базы данных, сравниваются с информацией масс-спектра, чтобы получить ранжированный список из 500 наиболее подходящих последовательностей (Шаг 3). Затем эти 500 последовательностей подвергаются корреляционному анализу для получения окончательной оценки и ранжирования последовательностей (Шаг 4) [3, с. 979].

Первый элемент программы включает предварительный поисковый анализ и обработку данных тандемной масс-спектрометрии. Отношения массы иона к заряду преобразуются в округлённые номинальные значения (ближайшее целое число). Чтобы сопоставить тандемный масс-спектр с последовательностью в базе данных, белковые последовательности извлекаются и сканируются, чтобы найти линейные комбинации аминокислот, идущие от *N* к *C*-концу, которые соответствуют массе пептида. Массы аминокислотных последовательностей

суммируют до тех пор, пока масса пептида не попадёт в заданный допуск по массе. Как правило, этот допуск устанавливается равным & 0,05 % или минимум 1 u, хотя это значение можно отрегулировать. По мере увеличения допуска по массе количество совпадающих последовательностей увеличивается.

Массу к заряду ионов-фрагментов, которые теоретически соответствуют аминокислотным последовательностям, рассчитывают следующим образом (1):

$$b_n = \sum a_n + 1$$
$$y_n = MW - \sum a_n \quad (1)$$

где a_n – масса аминокислоты, b_n – ион типа b, y_n – ион типа y. Эти значения используются на третьем этапе для сравнения последовательностей со списком отношения массы к заряду, полученным из тандемного масс-спектра.

Химические модификации аминокислот могут учитываться при этом поиске путём изменения масс аминокислот, используемых для расчёта масс пептидов. Модифицированная аминокислота затем рассматривается при каждом появлении в последовательности.

Метод оценки. Как только аминокислотная последовательность соответствует определённому допуску по массе, последовательность оценивается с использованием нескольких различных критериев. Во-первых, суммируются количество (n_i) предсказанных ионов-фрагментов, которые совпадают с ионами, наблюдаемыми в спектре в пределах ± 1 u, и их плотности (i_m). Непрерывность ряда ионов рассматривается путём увеличения компонента оценки (β) для каждого последовательного совпавшего фрагмента иона. Если ион иммония для аммокислот His, Tyr, Trp, Met и Phe присутствует в спектре вместе с ассоциированной аминокислотой, то увеличивается дополнительный компонент оценки (ρ). Если аминокислоты нет в последовательности, то ρ уменьшается. Значения, используемые для β и ρ, составляют 0,075 и 0,15 соответственно. Общее количество предсказанных ионов последовательности также отмечено (n_t).

Оценка (2) рассчитывается для каждой аминокислотной последовательности с использованием отношения

$$S_p = (\sum i_m) n_i (1 + \beta)(1 + \rho)/n_t \quad (2)$$

Однако SEQUEST не сравнивает необработанные спектры с прогнозами. Вместо этого он делит спектр на 10 бинов и нормализует каждый к наиболее интенсивному пику в бине, эффективно удаляя относительную интенсивность ионов по всему спектру фрагментации как важный фактор, определяющий совпадение. Данный метод оказался эффективным для соотнесения масс-спектров с потенциальными последовательностями даже без точных правил для предсказания интенсивности ионов фрагментов.

SEQUEST является очень точным и надёжным алгоритмом для идентификации белков в протеомике, однако он имеет некоторые недостатки. Во-первых, алгоритм SEQUEST работает только с фиксированными модификациями аминокислотных остатков. Это ограничение может привести к проблемам при идентификации неожиданных или неизвестных модификаций белковых последовательностей. Во-вторых, SEQUEST чувствителен к качеству экспериментальных данных. Некачественные масс-спектры могут привести к неверным результатам идентификации белков. Наконец, SEQUEST является времязатратным алгоритмом, который требует мощных вычислительных ресурсов.

MASCOT [12, с. 3176] – это алгоритм, который включает в себя несколько стратегий поиска в базе данных, используя данные тандемной масс-спектрометрии, такие как ионный поиск MC/MC и запросы по последовательностям. Он рассчитывает теоретические массы фрагментных ионов, что является общим подходом для многих алгоритмов поиска в базе данных, и затем сравнивает их с экспериментальными спектрами. Sequence Query предполагает частичную ручную интерпретацию данных MC/MC, которая включает определение молекулярной массы и характеристик последовательностей-кандидатов. Стратегии, используемые в MASCOT, основываются на вероятностной оценке с применением алгоритма MOWSE, учитывающего распределение размеров пептидов по отношению к массам белков в базе данных для их идентификации. Фундаментальный подход заключается в вычислении вероятности того, что наблюдаемое совпадение между набором экспериментальных данных и каждой записью в базе данных последовательностей является случайным событием. Совпадение с наименьшей вероятностью считается лучшим совпадением. Является ли наилучшее совпадение также значимым совпадением, зависит от размера базы данных. Для каждого поиска даётся пороговое значение вероятности того, что совпадение является чисто случайным событием.

X!Tandem [4, с. 1466][50] является самым популярным алгоритмом с открытым исходным кодом, оптимизированным по скорости и предназначенным для работы на скромных вычислительных ресурсах. Алгоритм делает одно важное аксиоматическое предположение: для каждого идентифицируемого белка в исходной белковой смеси будет существовать по крайней мере один обнаруживаемый триптический пептид с нулем или одним пропущенным сайтом расщепления. Уточнённый или вторичный анализ предполагаемых записей белков затем более полно анализируется (аналогично устойчивым к ошибкам поискам) с учётом неспецифического гидролиза и/или ПТМ.

Текущая встроенная функция оценки X!Tandem вычисляет оценку на основе скалярного произведения между теоретическим (только ионы *b* и *y*) и экспериментальным тандемным масс-спектром. Оценка впоследствии преобразуется в ожидаемое значение (*E*-значение). *E*-значение представляет собой количество пептидов в базе данных, которые, как ожидается, достигнут этой оценки только случайно (случайное совпадение). Чем ниже *E*-значение, тем значительнее оценка. *E*-значение получается путём сбора статистики во время

поиска для оценки распределения баллов для случайных и ложных идентификаций. Предполагается, что это распределение является гипергеометрическим, то есть дискретным распределением вероятностей, и поэтому значение *E* для пептидов с высокими показателями может быть получено путём экстраполяции. Для оценки совпадения теоретического и экспериментального спектров масс-спектрометрии используется специальный рейтинг, который вычисляется на основе формулы (3):

$$z = m_c! \, m_d! \sum_{i=0}^{m} I_i P_i \tag{3}$$

где m_c и m_d обозначают количество зарегистрированных ионов *c* - серии и *d* - серии в экспериментальном масс-спектре соответственно; сумма $\sum_{i=0}^{m} I_i P_i$ представляет собой скалярное произведение векторов экспериментального и теоретического масс-спектров, которое используется для оценки степени совпадения между ними.

Для определения надёжности идентификации белка используют рейтинг белка E_{pro}, который рассчитывается на основе формулы (4), учитывая достоверность каждого пептидного спектра, связанного с данным белком:

$$E_{pro} = \sqrt{M} \prod_{i=1}^{m} e_{i(z_i^*)} \tag{4}$$

где *M* – представляет собой общее число спектров, а *m* – число спектров, которые можно связать с конкретным белком [51]. Программное обеспечение X!Tandem способно идентифицировать пептиды даже с неполным или неспецифическим расщеплением, а также обнаруживать пептиды с посттрансляционными модификациями в сжатые сроки.

Эта оценка аналогична току интенсивности ионов *c* и *d*, которая вычисляется путём суммирования интенсивностей всех обнаруженных ионов в экспериментальных спектрах. Однако это не то же самое, что и использование данных об интенсивности пиков, которые могут указывать на процессы, влияющие на химическую фрагментацию. Данный показатель лишь подтверждает факт присутствия пика в спектре. Статистический анализ в X!Tandem использует *E*-значение для оценки разницы между наилучшим и другими совпадениями последовательностей-кандидатов, что является ключевым показателем результатов. Однако, поскольку подобный подход применяется в различных алгоритмах, одного *E*-значения недостаточно для существенного улучшения результатов X!Tandem. Алгоритмы, превосходящие X!Tandem, обычно включают дополнительные этапы и информацию для более точного присвоения спектров пептидам.

Главное нововведение X!Tandem – проводить поиск в два этапа. На первом этапе быстрый обзор идентифицирует белки-кандидаты, которые приблизительно соответствуют входным спектрам. На этом этапе предполагается идеальное расщепление и не допускаются никакие посттрансляционные модификации. На втором этапе новый поиск проводится

только по кандидатам, идентифицированным на первом этапе, на этот раз допуская уточнения, такие как пропущенные расщепления и посттрансляционные модификации, что значительно увеличивает сложность поиска. Выполнение этого уточненного поиска по меньшему количеству кандидатов из первого этапа значительно сокращает время поиска.

OMSSA (Open Mass Spectrometry Search Algorithm) [5, с. 958] – ещё один пример алгоритма с открытым исходным кодом, который уникален тем, что в нём используется классическая проверка гипотез, основанная на явной модели сопоставления статистики, типе статистической модели, используемой в BLAST. OMSSA берёт экспериментальные спектры MC/MC, фильтрует шумовые пики, извлекает значения *m/z*, а затем сравнивает эти значения *m/z* с расчётными значениями *m/z*, полученными из пептидов, полученных путём расщепления *in silico* библиотеки белковых последовательностей. Теоретические пептиды должны иметь массу в пределах указанного пользователем допуска массы предшественника. Полученные результаты поиска затем статистически оцениваются. Сам алгоритм подсчёта очков использует явную математическую модель для вероятности совпадения. Предполагается, что количество совпадений между наблюдаемыми пиками и теоретическими фрагментными ионами для данной пептидной последовательности может быть описано распределением Пуассона со средним значением (лямбда), рассчитанным на основе фрагментарной устойчивости к ионам, числа теоретических и экспериментальных пиков, и нейтральной массы иона-предшественника. Дальнейшие уточнения требуют, чтобы по крайней мере один из предсказанных фрагментных ионов соответствовал одному из трёх (по умолчанию) наиболее интенсивных пиков в спектре. Полученная оценка представляет собой вероятность, при условии, что модель верна и совпадение между наблюдаемым и теоретическим спектром было получено случайно; более низкое значение указывает на лучшее совпадение. OMSSA сообщает результаты в соответствии со значением *E,* связанным с каждым пептидом-кандидатом.

Phenyx [52] – коммерческий программный пакет для идентификации белков методом масс-спектрометрии. Он состоит из двух основных частей: алгоритма подсчёта очков и клиентского интерфейса, который можно использовать для визуализации, проверки или сравнения результатов, полученных с помощью нескольких различных алгоритмов поиска. Алгоритм поиска основан на алгоритме вероятностной оценки Olav, первоначально разработанном в Geneprot [53]. Алгоритм, который вычисляет оценку для каждого совпадения между экспериментальным спектром и пептидом из базы данных последовательностей, основан на нескольких индивидуальных оценках. Эти индивидуальные оценки зависят от различных свойств рассматриваемого пептида, включая, например, наличие или отсутствие ионов конкретных фрагментов, наблюдение последовательных ионов в ряду фрагментов, наблюдаемую ошибку массы исходного вещества или фрагмента или наблюдаемую интенсивность пиков. Для каждой из этих оценок рассчитывается отношение правдоподобия, которое сравнивает вероятность получения оценки при условии, что совпадение правильное, с вероятностью получения оценки при

случайном совпадении. Затем различные подоценки объединяются путём сложения суммы логарифмов каждого отношения правдоподобия, предполагая, что различные компоненты независимы. Наконец, баллы нормализуются с использованием распределения, полученного для набора случайно выбранных пептидов. Этот случайный шаг подразумевает, что оценка для данного пептида и экспериментального спектра может варьироваться в течение нескольких экспериментов. Алгоритм оценки может быть адаптирован для данных, поступающих от разных приборов или экспериментальных условий (например, ряд ожидаемых наблюдаемых фрагментных ионов может различаться в зависимости от типа прибора, используемого для сбора данных). В программе предусмотрены различные методы оценки по умолчанию для наиболее распространенных инструментов, но алгоритм также может быть оптимизирован для ранее неизвестного инструмента.

Таким образом, этот метод анализа масс-спектрометрии интегрирует структурную информацию, такую как интенсивность пиков, их смежность в ионных рядах и соотношение сигнал/шум, помимо данных о массе/заряде (*m/z*). Расширенная оценка совпадения, которая включает эти параметры, предоставляет более точное отражение качества идентификации совпадения. При анализе известных спектров Phenyx оценивает вероятности совпадений, различая правильные и случайные совпадения, и на основе этого выдаёт оценку. В случае идентификации пептидов из неизвестных спектров используется аналогичный подход с расширенной информацией о совпадениях, которая может быть сгенерирована для последовательностей-кандидатов в данной базе данных для определения оценки отношения. Оценка позволит отличить истинные совпадения от ложных.

Алгоритм идентификации PEAKS DB стабильно набирает популярность, поскольку он сочетает в себе функции поиска по базе данных и секвенирования *de novo*. В частности, он выполняет поиск в базе данных и использует секвенирование *de novo* для проверки результатов поиска. Поскольку для секвенирования *de novo* база данных не требуется, маловероятно, что совпадение между результатом поиска в базе данных и последовательностью *de novo* будет случайным событием. Эта уникальная функция помогла PEAKS DB значительно повысить производительность по сравнению со многими другими алгоритмами поиска в базе данных.

Основные алгоритмические шаги PEAKS DB выполняются следующим образом:

1. Секвенирование *de novo*: алгоритм PEAKS используется для выполнения секвенирования *de novo* для каждого входного спектра.
2. Составление короткого списка белков: метки последовательностей *de novo* используются для поиска приблизительных совпадений в базе данных последовательностей белков. Все белки в базе данных оцениваются в соответствии с совпадениями тегов последовательностей. 7000 лучших вариантов белков составляют окончательный список белков и используются в будущем анализе.

3. Составление короткого списка пептидов: все пептиды из короткого списка белков используются для сопоставления спектров MC/MC с функцией быстрой оценки. Для каждого спектра MC/MC сохраняются только 512 пептидов-кандидатов с наивысшей оценкой (включая пептиды с ПТМ).

4. Оценка пептидов: из 512 кандидатов, рассчитанных на этапе составления короткого списка пептидов, используется точная функция оценки, чтобы найти лучший пептид для каждого спектра. Сходство между последовательностью *de novo* и пептидом базы данных является важным компонентом функции подсчёта очков. Кроме того, оценка нормализована, чтобы её можно было сравнивать по разным спектрам.

5. Подтверждение результатов: для определения минимального порога оценки совпадения спектра пептидов для удовлетворения требований FDR пользователя используется модифицированный подход-приманка.

6. Вывод белков и группировка: пептиды с высокой степенью достоверности, идентифицированные на вышеуказанных этапах, используются для вывода белков. Те белки, которые имеют один и тот же набор совпадений пептидов, сгруппированы вместе для более удобного отчёта.

MS-GFDB [15, с. 2841] – это недавно разработанный алгоритм поиска в базе данных. В MS-GFDB используется метод производящих функций (MS-GF), который вычисляет точные значения *p* совпадений пептидного спектра на основе гистограммы спектрально-специфических показателей всех пептидов. Значения *p* MS-GF зависят только в PSM (а не в базе данных), поэтому могут использоваться в качестве альтернативной функции оценки для поиска в базе данных. Программный инструмент MS-GFDB использует два типа метрик для оценки соответствия между спектром и потенциальным пептидом. Один из типов метрик – это оценка MS-GF, которая помогает оценить качество соответствия между спектром и пептидом. Вторая метрика – это *p*-значение, которое используется для определения статистической значимости соответствия. Для расчёта оценки MS-GF программа сначала преобразует исходный спектр в так называемый спектр массы префикса-остатка (PRM), что предполагает применение специальных параметров, выбранных на основании метода фрагментации и используемого фермента. Этот спектр PRM можно представить как адаптированную версию исходного спектра, содержащую набор оценок для каждой возможной массы аминокислотного остатка в пептиде вплоть до полной массы исходного пептида. Оценка для конкретной массы *m* в спектре PRM рассчитывается как логарифм отношения вероятностей того, что исследуемый пептид содержит аминокислоту с префиксной массой *m*. Следовательно, общий балл PSM получается путём суммирования всех оценок в спектре PRM, которые соответствуют префиксным массам анализируемого пептида. Этот процесс позволяет более точно оценить, насколько вероятно, что данный пептид действительно соответствует измеренному спектру. Чтобы вычислить значение *p*, MS-GF генерирует гистограмму оценок всех пептидов, используя подход с производящей функцией. Этот алгоритм даёт пользователю возможность создавать свои уникальные функции подсчёта очков (функции

генерации) для различных протоколов, что делает его более специфичным для различных видов экспериментов.

1.4 Исследование проблемы определения функций белков

Предсказание функции белка является серьёзной проблемой в области биоинформатики, целью которой является предсказание функций, выполняемых известным белком. Многие формы данных о белках, такие как последовательности белков, структуры белков, сети взаимодействия белок-белок и представления данных микрочипов, используются для прогнозирования их функций. Протеины состоят из длинных последовательностей аминокислот, синтезированных на основе инструкций из ДНК, и они сворачиваются в уникальные трёхмерные структуры. Трёхмерная конфигурация, которую принимает белок, диктуется кодом, вшитым в ДНК, и структура ДНК, в свою очередь, зависит от последовательности аминокислот. В биологии именно структура молекулы определяет её функцию.

В то время как такие методы, как микрочиповый анализ, РНК-интерференция и двухгибридная система дрожжей, могут использоваться для экспериментальной демонстрации функции белка, достижения в технологиях секвенирования сделали скорость, с которой белки могут быть экспериментально охарактеризованы, намного ниже, чем скорость, с которой становятся доступными новые последовательности [54]. Таким образом, аннотирование новых последовательностей в основном выполняется с помощью компьютерного предсказания, поскольку эти типы аннотаций часто можно выполнить быстро и для многих генов или белков одновременно. Первые такие методы предполагали функцию, основанную на гомологичных белках с известными функциями (предсказание функции на основе гомологии). Развитие контекстно-ориентированных и структурно-ориентированных методов расширило объём информации, которую можно предсказать, и теперь можно использовать комбинацию методов для получения картины полных клеточных путей из данных о последовательности. Важность и распространенность компьютерного прогнозирования функции генов подчёркивается анализом «доказательных кодов», используемых в базе данных Gene Ontology: 98% аннотаций были перечислены под кодом IEA (полученным из электронной аннотации), в то время как только <0,1 % из более чем 179 миллионов белков в UniProt были основаны на экспериментальных данных [55].

Функциональное предсказание затруднено, отчасти и потому, что природа «функции» плохо определена. Многие белки выполняют несколько функций, и почти в каждом семействе белков есть члены, имеющие очень разные функции. Более того, существуют десятки примеров негомологичных белков со схожими функциями. Известно, что гомологичные белки могут иметь разные функции, а негомологичные белки напротив могут иметь сходные функции. Прогнозы на более коротких эволюционных расстояниях, когда был аннотирован близкородственный гомолог, всегда будут более надёжными, в то время как методы, использующие наиболее чувствительные методы обнаружения гомологии, хотя и точны в отношении гомологии, вероятно, будут гораздо менее

точны в функциональном выводе. В целом, однако, сходные белки (белки, имеющие > 50 % идентичности) имеют сходные функции, и белки, идентичность которых составляет от 30 до 50 %, скорее всего, принадлежат к одному функциональному классу. Но для гораздо более отдаленных отношений вывод о гомологии может лишь немного улучшить предсказание функций.

Обычно функции белков определяются путём ручной или автоматизированной аннотации. Ручной процесс, проводимый специалистами, считается эталоном для аннотации функций из-за высокого качества результатов. Однако он требует значительных затрат и времени, что делает его малопригодным для использования в больших масштабах.

Также следует отметить, что благодаря развитию передовых технологий секвенирования нового поколения, известных как NGS (Next-Generation Sequencing), значительно увеличилось количество генетических последовательностей, которые требуют аннотации. Это привело к созданию методов вычислительной аннотации, обусловленных потребностью в автоматизации обработки масштабируемых объёмов недавно секвенированных последовательностей и стремлением к повышению точности аннотированных данных.

Ввиду обширного разнообразия терминологии для классификации функций белков, были созданы специализированные базы данных, направленные на унификацию этой классификации. К таким базам данных относятся Киотская энциклопедия генов и геномов (KEGG), Комиссия по ферментам (Enzyme Commission, EC), а также Функциональный каталог (FunCat).

На данный момент проект Gene Ontology (GO) считается наиболее универсальным и полным инструментом, так как он включает все необходимые атрибуты для системы функциональной классификации. Консорциум Gene Ontology разработал базу данных с управляемым словарём для описания функциональных характеристик продуктов генома, включая гены, белки и РНК.

Каждый набор терминов в Gene Ontology, называемый онтологией, относится к одной из трёх основных категорий: молекулярные функции (МФ, MF, Molecular Function), биологические процессы (БП, BP, Biological Process), и клеточные компоненты (КК, CC, Cellular Component). Gene Ontology представляет собой иерархическую систему, структурированную как направленный ациклический граф (DAG), где каждый термин является узлом, связанным с другими узлами отношениями типа «является» («is-a») или «является частью» («part-of»). Эта структура позволяет извлекать различные уровни информации и обеспечивает гибкость в аннотации функций белков, давая возможность маркировать их в широком спектре от общих до очень специфичных функций на основе имеющихся данных.

Автоматизированное прогнозирование функций (Automated Function Prediction – AFP) на основе системы Gene Ontology является одной из наиболее сложных областей биоинформатики. В ряде исследований поднимаются вопросы, связанные с присвоением функций белкам, исследуя их с различных научных позиций. Ранее публиковались обзоры, целиком посвященные

вопросам AFP [56], особенно с учётом типов данных, используемых в исследованиях. Эти обзоры не только касались различных методов и подходов к автоматизированному прогнозированию, но и анализировали эффективность различных стратегий сбора данных [57], их обработки и последующего использования в функциональной аннотации, недостатков и соответствующих решений, сети взаимодействия белков, типов классифицированных функций и назначения терминов Gene Ontology на основе информации о последовательности. Они выделяют важность выбора подходящего типа данных для улучшения точности и надёжности предсказания функций белков.

В [58] демонстрируется предсказание функции белка в рабочем процессе машинного обучения. В [59] рассматривается прогнозирование функции генов с точки зрения моделирования Gene Ontology. Все эти исследования представили независимые взгляды на проблему, однако не представлено подробного обзора методов глубинного обучения, которые является новым подходом к прогнозированию функций белков с помощью терминов Gene Ontology.

Быстрорастущей областью исследований является применение машинного обучения для классификации белков. Общая схема решений машинного обучения, детали которых могут различаться в зависимости от метода, показана на рисунке 2 [60].

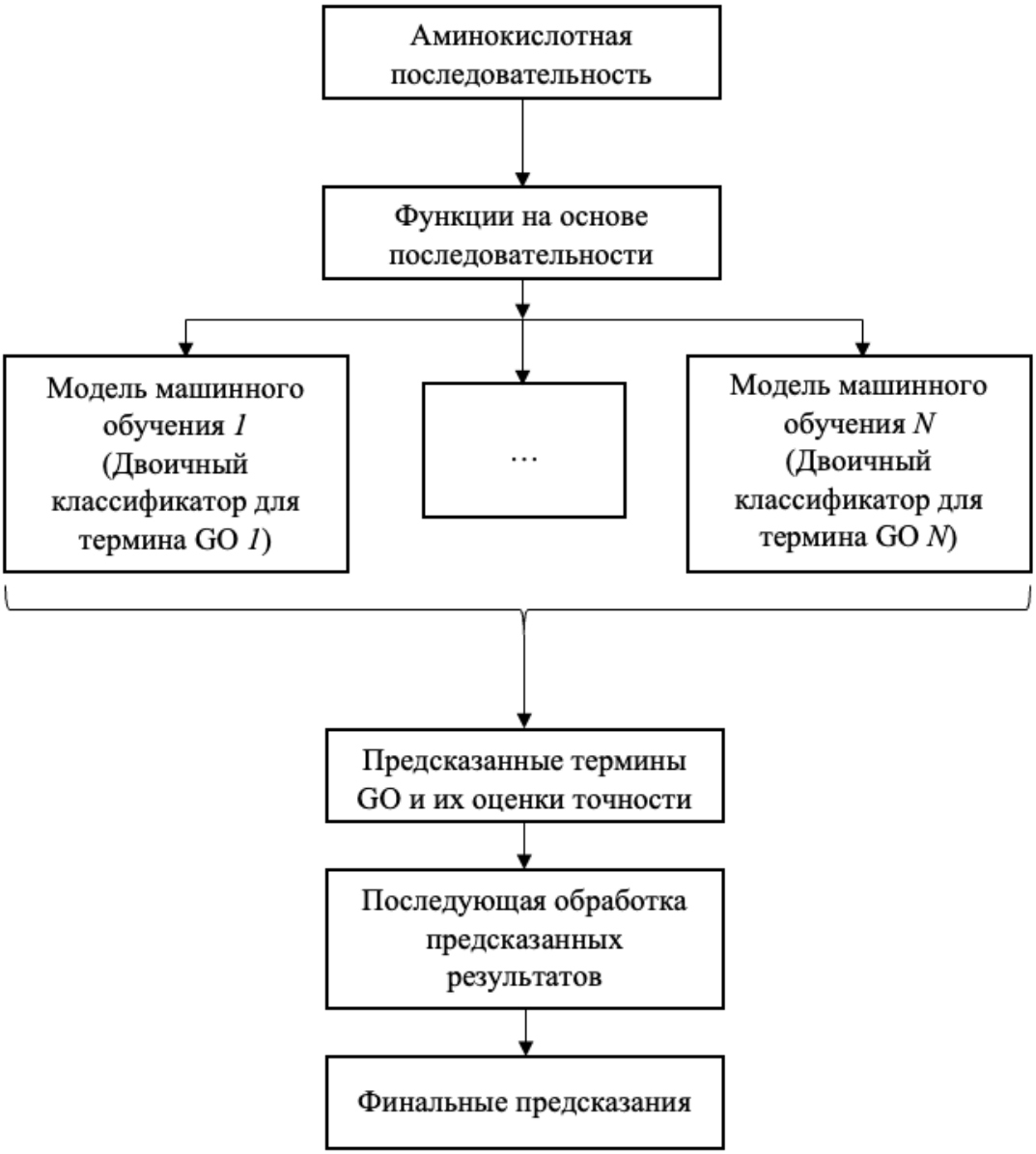


Рисунок 2 – Общий алгоритм решений машинного обучения для прогнозирования терминов GO белков

Нейронные сети являются наиболее широко используемыми инструментами машинного обучения, применяемыми для решения задач классификации. Прогнозирование функции белков также можно рассматривать как проблему классификации.

Прогнозирование функции белка представляет собой задачу мультиклассовой классификации, где имеется набор возможных функций $F=(F_1,\ldots,F_m)$. Для данного списка белков $P=(P_1,\ldots,P_n)$, ситуация такова, что первые l белков уже классифицированы со значениями $y_1,\ldots,y_l$. Каждый элемент y_i – это вектор, в котором $y_{ij}=1$, если белок P_i выполняет функцию F_j, и $y_{ij}=0$ в противном случае. Задача состоит в определении функций $y_{1+1},\ldots,y_n$ для оставшихся белков $P_{1+1},\ldots,P_n$, которые ещё не были классифицированы [61]. Это требует определения связи между каждым белком и функциями, которые он

может выполнять, основываясь на уже известной информации и предсказаниях для тех белков, для которых данные отсутствуют.

Такая формулировка подразумевает необходимость применения машинного обучения и разработки моделей, которые обеспечат оптимальное сочетание точности классификации и скорости вычислений. Преимуществом машинного обучения является его способность к улучшению производительности и точности с ростом объёма тренировочных данных. С учётом растущего количества доступных белковых последовательностей, машинное обучение становится особенно перспективным для автоматизированного прогнозирования функций белков, что делает его актуальным направлением научных исследований.

Для решения задач классификации объектов разработано огромное количество нейронных сетей. Это открывает большие возможности для их использования для задачи предсказания функций белков, рассматриваемой как задача классификации. Однако на основании статистических данных, демонстрирующих, что большинство последовательностей белков в базах данных ещё не получили статус охарактеризованных, можно утверждать, что возможности нейронных сетей недостаточно изучены в плане функциональной аннотации белков. Это приводит к целесообразности изучения этих нейронных сетей и их многократного тестирования на разных наборах данных, что поможет определить их сильные и слабые стороны, модифицировать и повысить их точность для решения поставленной выше задачи. Поэтому исследования, которые посвящены изучению моделей машинного обучения и оценке их точности для решения задачи прогнозирования функций белков, имеют научную актуальность. Основываясь на рассмотрении вышеперечисленных проблем, очень важно разработать новый метод прогнозирования для решения проблемы прогнозирования функции белка.

1.5 Анализ существующих алгоритмов и решений для предсказания функций белковых последовательностей

За последние несколько десятилетий было получено множество данных о последовательностях белков с использованием высокопроизводительных методов, что делает их подходящим кандидатом для прогнозирования функций белков с использованием методов глубокого обучения. До сих пор было предложено много таких передовых методов. В этом разделе представлены всесторонние сведения о новейших методологиях, их плюсах и минусах, а также о точности прогнозирования и новом направлении с точки зрения интерпретируемости прогностических моделей, которые необходимо использовать в системах прогнозирования функций белков.

Изначально аннотации для неописанных белков назначались, опираясь на базовый метод: поиск белковой последовательности в базах данных, содержащих экспериментально подтверждённые белки. Если находились белки, схожие по определённым критериям, связанные с ними термины Gene Ontology передавались искомому белку. Ранние методы определения функциональности на основе сходства были построены на принципах гомологии, и использовались

инструменты для локального выравнивания последовательностей, например, BLAST. При использовании таких методов неизвестную белковую последовательность сопоставляется с последовательностями из базы данных, содержащей хорошо аннотированные белки с уже известными функциями. Затем идентифицируется извлечённая последовательность с наивысшей оценкой выравнивания в соответствии с заданным порогом, и её аннотация передаётся запрашивающему белку. OntoBlast, GOFigure, GOblet и GOtcha являются типичными системами аннотаций, использующими сходство последовательностей, определяемое поисковым механизмом BLAST. Более подробная информация об этих инструментах представлена в таблице 1.

Таблица 1 – Инструменты для аннотирования белковых последовательностей терминами GO, основывающиеся на традиционном подходе

Тип подхода	Наименование алгоритма или программного модуля	Описание	Год
1	2	3	4
Подходы, основанные на сходстве последовательностей	OntoBlast	Онлайн-инструмент, являющийся подмодулем инструмента «Ontologies TO GenomeMatrix». Работает по классическому алгоритму предсказания функций белков с использованием взвешенного списка терминов Генной Онтологии, которые связаны с последовательностями, ранее аннотированными при помощи девяти биологических баз данных.	2003
	GOFigure	Производит вывод в виде кликабельного графа в четыре этапа, включая поиск гомологичной последовательности, построение графа с минимальным покрытием и назначение онтологий после их оценки.	2003
	GOblet	Программный модуль, в основе которого лежит метод определения чувствительности к *E*-значению и список баз данных, работающих с BLAST. После идентификации белков-кандидатов на основе найденных онтологий строится направленный граф, в котором количество последовательностей, имеющих общий термин GO, является кумулятивным, чтобы представить значимость термина.	2003
	Gotcha	Переносит ассоциации Gene Ontology, полученные из BLAST, в генные продукты с помощью новой схемы ранжирования. Как и в предыдущем инструменте на основе идентификаций строится направленный ациклический граф, в котором *E*-значения	2004

Продолжение таблицы 1

1	2	3	4
		родительских терминов учитываются для их производных терминов. Это даёт возможность получить нормализованные оценки достоверности для каждой из трёх подонтологий Gene Ontology.	
	PFP	Программа для функционального аннотирования белков, в основе которой лежат совпадения PSI-BLAST. Метод использует совпадения для последовательностей, которые достигают только очень высокого порога точности, что позволяет более точно извлекать подходящие термины GO из базы данных UniProt и присваивать их неизвестным белкам.	2009
	INGA	Ещё один онлайн-инструмент функционального аннотирования. Работает с большим количеством биологических баз данных, присваивая термины GO неизвестным белкам.	2015
	GoFDR	Инструмент, в основе которого лежит метод множественного выравнивания с использованием оценки ложных обнаружений (false discovery rates) и матрицей оценок для конкретной позиции (position-specific scoring matrix) для сортировки терминов Генной Онтологии для каждой аминокислотной последовательности.	2016
Вероятностные подходы	PFP from PPI	Вероятностный алгоритм, основанный на сетях белок-белковых взаимодействий для предсказания терминов Gene Ontology. Сеть была протестирована на специфичных данных неаннотированных белков в дрожжах.	2003
	ProbPFP	Инструмент для функционального аннотирования белков *Saccharomyces Cerevisiae*. Использует дополнительные данные для повышения точности вероятностной модели.	2007
	BMRF	Вариант подхода на основе MRF к Protein-Protein Interaction, который позволяет делать новые прогнозы для неаннотированных белков.	2010
Подходы, основанные на машинном обучении	GOPET	Инструмент прогнозирования и оценки терминов GO, предоставляющий термины GO для молекулярных функций и биологических процессов для любого организма.	2006

Продолжение таблицы 1

1	2	3	4
	PoGO	Модель ансамбля со сходством последовательностей InterPro, биохимическими свойствами и третичной структурой белка для прогнозирования GO белков грибов.	2010
	FFPred3	Инструмент на основе SVM, обеспечивающий независимое от гомологии назначение терминов GO для эукариотических белков.	2016
	PANNZER2	Взвешенный классификатор *k*-ближайших соседей, обеспечивающий функциональную аннотацию для неохарактеризованных белков с терминами GO и описаниями в произвольном тексте.	2018
	DeepText2GO	Инструмент прогнозирования функций белков, комбинирующий метод текстового поиска и метод, основанный на сходстве последовательностей.	2018
	NetGo	Онлайн-инструмент, сочетающий использование сетей белок-белковых взаимодействий и сходства последовательностей для присвоения терминов Генной Онтологии неохарактеризованным белкам.	2019

PFP представляет собой подход к функциональному аннотированию, который учитывает информацию о белках, гомологичных исследуемым, с использованием PSI-BLAST. Этот инструмент является улучшенной версией в плане расширения охвата и повышения точности аннотаций, что было подтверждено анализом на стандартных наборах данных в сравнении с другими утилитами, такими как Gotcha и InterProScan. Также существует метод, известный как INGA [18, с. 2], который объединяет данные из сетей белок-белковых взаимодействий с информацией о доменах и сходствах последовательностей, полученных из BLAST. Этот метод способствует синтезу широкого спектра данных для достижения более обоснованных прогнозов функций белков согласно Gene Ontology, применяя анализ обогащения. GoFDR [19, с. 4], в свою очередь, экстраполирует релевантные GO термины из запросов, проведённых через систему множественного выравнивания последовательностей (Multiple Sequence Alignment – MSA), используя для этого BLAST или PSI-BLAST. Этот подход позволяет включать широкий диапазон белковых последовательностей в поиск соответствующих функциональных аннотаций, что способствует более всестороннему и точному функциональному аннотированию. Вероятность присвоения термина последовательности запроса определяется функционально различающими остатками (Functionally Discriminating Residues), оценочной матрицей для конкретных позиций для

функционально различающихся остатков и таблицей оценки с максимальной вероятностью, основанной на наборе обучающих последовательностей.

Методы, основанные на локальном выравнивании последовательностей, сохраняют свою значимость в биоинформатике благодаря своей простоте и способности достигать надёжных результатов во многих ситуациях. Однако они не лишены определённых ограничений, таких как потенциальные ошибки из-за некорректных аннотаций в базах данных или избыточной передачи функций между негомологичными последовательностями. Дополнительные сложности включают в себя выбор соответствующих пороговых значений для выявления значимых выравниваний, что может повлиять на чувствительность и специфичность результатов.

Кроме того, в арсенале современных исследователей имеются альтернативные предикторы, которые используют более продвинутые подходы, такие как сравнение сходства структур белков, определение принадлежности к определённым белковым семействам или применение филогеномических методов. Эти подходы позволяют предсказывать функции белков, опираясь на более глубокий анализ структурных характеристик или исторического развития белковых семейств, что может обеспечить более точную и целостную аннотацию, особенно в случаях, когда прямое сравнение последовательностей не даёт достаточно информации для надёжной аннотации.

В [62] была разработана серия вероятностных моделей, предназначенных для определения функций белков на основе графа функциональных связей, построенного из данных о белок-белковых взаимодействиях дрожжей *Saccharomyces cerevisiae*. Основная гипотеза исследования заключалась в том, что белки, которые тесно связаны в сети PPI, вероятнее всего разделяют схожие функции, что отражается на вероятностном распределении терминов Gene Ontology. Для оценки этих вероятностей использовался метод, который опирался на биномиальную модель и включал в себя алгоритм случайного поля Маркова (MRF). Этот подход позволил учесть комплексные взаимодействия между белками, существующие в сети PPI, и тем самым повысить точность присваиваемых функциональных аннотаций. Со временем методика была расширена за счёт интеграции дополнительных источников данных, включая информацию о генной экспрессии, белковых мотивах, мутантных фенотипах и локализации белков. Использование байесовского подхода для объединения этих разнообразных данных позволило создать мощную статистическую модель, которая превзошла предыдущие модели, опирающиеся только на данные PPI. Этот комплексный подход улучшил точность и надёжность функционального прогнозирования, давая более полное представление о функциональных возможностях белков в различных биологических контекстах.

В рамках исследований, фокусирующихся на дрожжах *Saccharomyces cerevisiae*, была предложена байесовская интерпретация модели марковских случайных полей в работе [63], что представляет собой продвижение в оценке параметров модели и обеспечении более точного прогнозирования, опираясь на данные сетевых взаимодействий. Этот подход учитывает предыдущее знание и интегрирует его с наблюдаемыми данными для улучшения статистических

выводов. Дополнительно, в [64] были оценены различные методы взвешивания в контексте трёх алгоритмов: tSVD (усечённое сингулярное разложение), SIM (семантически улучшенное tSVD) и pLSAnorm (вероятностный латентно-семантический анализ с нормализацией).

Инструментарий машинного обучения был спроектирован для обнаружения неявных корреляций между многообразными атрибутами белков, включая их аминокислотную последовательность, трёхмерную структуру, а также различные эволюционные маркеры, и их функциональными характеристиками, определёнными через термины Gene Ontology. Эти системы обучаются на основе данных, состоящих из полноценно описанных макромолекул, и применяют полученные знания для аннотации белков, которые ещё не были исследованы. Разработанные методики машинного обучения агрегируют и анализируют обширные биологические и биоинформатические датасеты, их структурные, последовательностные и эволюционные сигналы для точного предсказания функций белка. Эти методы не только включают стандартные подходы к классификации, но и применяют передовые алгоритмы, способные учитывать комплексные и высокоразмерные данные, что позволяет с большей вероятностью предсказать широкий спектр функциональных активностей белков. Кроме того, современные подходы к машинному обучению могут интегрировать многослойные сети и методы глубокого обучения, такие как свёрточные нейронные сети и рекуррентные нейронные сети, для обработки и интерпретации последовательностных данных в сочетании с дополнительными биологическими инсайтами, такими как филогенетические деревья и паттерны мутаций.

Множественные машины опорных векторов (Support Vector Machine – SVM) являются предпочтительным инструментом в ряде научных работ, направленных на анализ и классификацию функций белков. Примером такого применения является инструмент GOPET [65], который интегрирует разнообразные функциональные сигналы, ассоциированные с терминами Gene Ontology, включая меры сходства последовательностей на основе BLAST, частоты, качество аннотаций гомологичных белков и их уровень в иерархии GO, для обучения набора из 99 SVM-классификаторов. Эти классификаторы должны определить, является ли конкретный термин GO «правильным» или «неправильным» для аннотации неизвестной белковой последовательности, при этом используются показатели достоверности, основанные на системе голосования. FFPred [66, 67] был изначально разработан для аннотации неописанного протеома человека, но затем его обобщили и адаптировали для применения к другим протеомам. Последнее воплощение этого инструмента, FFPred3 [20, с. 2], сохраняет основу на SVM, но было расширено для более глубокого изучения взаимосвязей между признаками, извлечёнными из последовательностей и структур, в контексте трёх основных подонтологий GO. Это расширение включает в себя более сложный анализ данных, который позволяет улучшить точность предсказаний и даёт более полное представление о функциональных аспектах белков.

Prediction of Gene Ontology terms (PoGO) [21, с. 2] ещё одна программа для аннотации функций белка, расширяющая возможности классификации путём интеграции нескольких типов данных. В отличие от методов, ограниченных использованием терминологии InterPro, PoGO сочетает в себе данные о сходстве последовательностей, биохимических свойствах и третичной структуре белков, что позволяет создавать более глубокие аннотации [68]. В процессе первоначальной классификации PoGO использует методы, основанные на опорных векторах и линейной классификации. Эти методы обеспечивают предварительное разделение данных на категории, которое затем усовершенствуется с помощью метаобучения. Метаобучение, или ансамблевое обучение, в данном контексте представляет собой процесс, при котором результаты нескольких моделей обучения агрегируются для улучшения точности предсказания окончательной классификации. Этот многоуровневый подход улучшает способность системы различать и классифицировать функциональные аспекты белков с большей точностью, исходя из сложного набора признаков, что делает PoGO весьма эффективным в контексте функциональной геномики.

Инструмент PANNZER2 [69] интегрирует методику *k*-ближайших соседей (*k*-Nearest Neighbor – KNN) для предоставления эффективных решений в функциональной аннотации белков. Этот подход использует гомологию последовательностей вместе с рядом других аннотационных предикторов для ускорения процесса аннотации. Преимущество KNN в этом контексте заключается в его способности улавливать и анализировать сложные многоуровневые отношения в больших биологических наборах данных, что делает его особенно полезным для идентификации функциональных сходств между белками на основе их последовательности и других значимых биоинформационных признаков. Между тем, MS-KNN объединяет разнородные данные, чтобы предложить конкурентоспособную модель для предсказания функции белка. DeepText2GO [22, с. 2] представляет собой интегрированный подход, который совмещает продвинутое глубокое семантическое понимание текста, полученного из цитат базы данных MEDLINE, с информацией о последовательности, извлечённой с помощью инструментов BLAST и InterProScan. Из этого сочетания текстовых и последовательностных данных формируются признаки, которые затем используются в таких машинно-обучающих моделях, как метод *k*-ближайших соседей и логистическая регрессия. Эти модели способны эффективно предсказывать функции белков, работая с большим объёмом данных и без предварительных знаний о функциях конкретных белков. NetGO [23, с. 380] является продолжением работы GoLabeler [70] и использует стратегию обучения ранжированию (Learning-To-Rank – LTR), что позволяет интегрировать информацию, основанную на последовательностях, для повышения точности функциональной аннотации белков. Преимущество NetGO заключается в использовании экстенсивной сети взаимодействий между белками, охватывающей более 2000 видов, благодаря доступу к базе данных STRING [71].

Методы машинного обучения выступают в качестве перспективного и растущего направления в области автоматического функционального предсказания. Это особенно важно для белков, которые либо являются новооткрытыми и не имеют установленных гомологий в доступных базах данных, либо для тех, чьи потенциальные гомологи ещё не ассоциированы с функциональными данными GO. Следствием этого является необходимость разработки методов, способных предсказывать функции белка "с нуля", исходя исключительно из первичной структуры аминокислотной последовательности, минуя традиционные подходы, зависящие от гомологичных сравнений или аннотированных библиотек. Такой подход требует создания точных и надёжных машинно-обучаемых моделей, которые могут обнаруживать сложные шаблоны и зависимости непосредственно в последовательности белка, что позволяет определить его потенциальную биологическую роль без необходимости опираться на уже существующие базы данных.

Технологии глубокого обучения продемонстрировали свой выдающийся потенциал в различных сферах, в том числе и в биоинформатике [72]. Уникальность этих методов заключается в их способности к самостоятельному обучению, в ходе которого они автоматически выделяют важные признаки из первичных данных, что позволяет строить сложные прогностические модели. В отличие от традиционных методов машинного обучения, которые требуют ручного создания и отбора признаков, глубокое обучение достигает высокой точности классификации благодаря изучению иерархии признаков напрямую из данных. С ростом объёмов генерируемых геномных данных увеличивается и потребность в сложных алгоритмах и вычислительных мощностях, что стимулирует дальнейшее развитие и внедрение глубокого обучения в задачи функциональной аннотации. Методы, предложенные в последних исследованиях, основываются на комплексной оценке моделей и разнообразии используемых данных, что позволяет более глубоко исследовать и понимать функциональную роль генов и белков. В таблице 2 представлен свод инструментов, использующих глубокое обучение для функционального аннотирования белков.

Таблица 2 – Методы, основанные на глубинном обучении, для присвоения белкам соответствующих терминов GO

Используемая функция	Название программы	Описание	Модель DL
1	2	3	4
Методы, основанные на последовательности	Deep autoencoder	Метод использует глубокий autoencoder для прогнозирования терминов Генной Онтологии	Autoencoder
	PFP_DRBM	Инструмент работает с данными *Homo sapiens*, *S. cerevisiae*, *Mus musculus* и *Drosophila*. В основе заложен алгоритм машин Больцмана с глубоким	Deep Restricted Boltzmann Machines

Продолжение таблицы 2

1	2	3	4
		ограничением для присвоения белкам терминов Gene Ontology	
	ProLanGO	В основе модели заложен алгоритм Long-short term memory, использующий нейронный машинный перевод для прогнозирования функций белков на основе их аминокислотных последовательностей.	Long-Short Term Memory
	SECLEF	Онлайн-инструмент, позволяющий проводить не только функциональное аннотирование, но также определять семейство белка, обращаясь в биологическую базу данных UniProt. Основной входящий параметр – аминокислотная последовательность.	Свёрточная сеть
	DEEPred	Инструмент для функционального аннотирования белков терминами Gene Ontology, использует набор связанных многозадачных глубоких нейронных сетей.	Глубокая сеть
	DeepSeq	Алгоритм прогнозирует функцию белковых последовательностей *Homo sapiens*. В основе заложен подход глубинного обучения с использованием свёрточной нейронной сети, использующей только мотивы последовательности.	Свёрточная сеть
	DeepGOPlus	Инструмент является расширенной и оптимизированной версией инструмента DeepGO. В отличие от предшественника использует свёрточные нейронные сети.	Свёрточная сеть
Методы, основанные на данных или структуре	PFP_CNN	Инструмент использует глубокую свёрточную нейронную сеть для классификации функций белков *Homo sapiens* на основе их третичной структуры.	Свёрточная сеть
	DeepGO	Онлайн-инструмент, использующий архитектуру глубокой свёрточной нейронной сети для функционального аннотирования с помощью проекта Gene Ontology. Иерархически структурированный классификатор выводит термины GO для каждого белка.	Свёрточная сеть
	deepNF	Инструмент, основанный на глубинном слиянии сетей, автоматически прогнозирует функции, используя сетевые данные.	Autoencoder

Продолжение таблицы 2

1	2	3	4
	PFP_MTDNN	Метод использует многослойные глубокие нейронные сети для функционального аннотирования белков на основе их последовательностей и структуры.	Глубокая сеть
	DeepFunc	Выводит аннотации Gene Ontology путём объединения функций сети на основе InterPro и взаимодействия белков.	Полносвязная глубока сеть
	DeepGOA	Интегрирует исчерпывающую информацию о последовательности и белок-белковых взаимодействиях для автоматизированного функционального аннотирования.	Двунаправленная Long-Short Term Memory, свёрточная сеть
	SDN2GO	Инструмент, интегрирующий архитектуры с тремя подмоделями и классификатором взвешивания для прогнозирования терминов Генной Онтологии.	Свёрточная сеть
	DeepAdd	Инструмент со структурой на основе свёрточной нейронной сети, предсказывающей функцию белка на основе последовательности и дополнительной информации о белок-белковых взаимодействиях или вторичной структуре.	Свёрточная сеть
	FFPred-GAN	Инструмент для функционального аннотирования, использующий обучающий набор данных, дополненный сетью глубокого обучения.	Генеративно-состязательная нейросеть
	GONET	Инструмент, в основе которого находится рекуррентная нейронная сеть для предсказания функций белка, использующий обучение представлению для встраивания аминокислотных последовательностей.	Свёрточная сеть, рекуррентная сеть

Обучение с учителем является важным подходом для прогнозирования функции белков *in silico* и играют ключевую роль в аннотировании функций белков в целом. Эти модели обучаются на основе внушительных наборов данных, состоящих из белков, которые были детально исследованы, а их функции подтверждены экспериментальными методами. Такой подход гарантирует, что данные для обучения содержат надёжную и точную информацию, что критически важно для обеспечения валидности предсказаний модели. В ходе обучения алгоритмы машинного обучения выявляют сложные

шаблоны и закономерности, связывающие специфические последовательности аминокислот с соответствующими молекулярными функциями.

После завершении обучения модель способна распознавать эти шаблоны в новых, ранее неизвестных белковых последовательностях и назначать соответствующие функциональные аннотации согласно системе классификации Gene Ontology. Это особенно полезно в условиях, когда количество новооткрытых белков растёт экспоненциально, превышая возможности традиционных экспериментальных методов аннотирования. Обученная модель может эффективно масштабироваться для работы с большими наборами данных, предоставляя ценные предварительные данные, которые могут направлять будущие эксперименты и исследования.

Кроме того, современные подходы в области машинного обучения исследуют возможность интеграции различных типов биологических данных, таких как генетические, транскрипционные и посттрансляционные модификации, для создания более всесторонних и мультифакторных моделей, которые могут предсказывать функцию белка с ещё большей точностью. Таким образом, обученные модели не только обеспечивают ценные предсказания для новых последовательностей, но и способствуют глубокому пониманию функциональной сложности белков в живых системах, что в свою очередь способствует прогрессу в медицинских и биотехнологических исследованиях.

Архитектура глубокой нейронной сети (Deep Neural Network – DNN) организована таким образом, что она состоит из последовательности слоёв: входного, нескольких промежуточных скрытых слоёв, каждый из которых выступает в роли трансформатора информации, и выходного слоя, который производит конечный результат обработки. Данные поступают на вход сети и последовательно проходят через каждый слой, где происходит их трансформация благодаря нейронным связям и функциям активации. Эта последовательная обработка информации позволяет DNN выполнить сложное прогнозирование или классификацию.

В контексте предсказания функций белков на основе Gene Ontology, глубокие нейронные сети могут применяться в качестве многозадачных систем глубоких нейронных сетей (MTDNN) [73], что позволяет одновременно решать несколько задач аннотации, обучаясь предсказывать множество функций белка сразу. Это повышает эффективность обучения, поскольку сеть учится распознавать разнообразные аспекты белковых функций, что увеличивает вероятность обнаружения глубоких и сложных закономерностей в биологических данных. Это улучшает точность прогнозируемых аннотаций, делая процесс более эффективным и надёжным

Свёрточные нейронные сети (Convolutional Neural Network – CNN) были спроектированы изначально как инструмент для анализа двумерных визуальных данных, особенно для идентификации и классификации рукописных чисел. Однако со временем их применение расширилось, и они доказали свою эффективность в обработке не только многомерных изображений, но также одномерных данных, таких как аудио, тексты и биологические последовательности, включая ДНК и белки.

Ключевой элемент CNN – свёрточный слой, в котором признаки входных данных извлекаются с помощью фильтров или ядер, проходящих по данным и создающих карты признаков. Эти фильтры эффективно сканируют данные, выделяя важные характеристики без необходимости понимания их положения в пространстве. Последующие слои пулинга (pooling layers) служат для уменьшения размерности карт признаков, сжимая информацию и оставляя только самое существенное, что позволяет уменьшить количество параметров и вычислительные затраты, а также увеличить устойчивость сети к незначительным изменениям во входных данных. В результате эти слои создают иерархию признаков, где низкоуровневые признаки в начальных слоях постепенно объединяются в слоях более высокого уровня, которые способны распознавать всё более сложные и абстрактные паттерны. Эта иерархия позволяет CNN распознавать объекты и особенности на различных масштабах и уровнях абстракции, что делает её мощным инструментом для изучения и классификации многомерных данных.

С течением времени архитектуры CNN были адаптированы для выполнения задач автоматического предсказания функций белка, где они могут использоваться самостоятельно для изучения и классификации биологических последовательностей [74, 75] или в комбинации с другими моделями машинного обучения [76]. В контексте биоинформатики это даёт возможность учёным анализировать и аннотировать белки и гены с высокой точностью, опираясь на комплексные паттерны в геномных данных. Комбинация свёрточных и пулинговых слоёв в CNN позволяет архитектуре обрабатывать большие и сложные наборы данных, такие как геномные последовательности, эффективно находя важные биологические сигналы в них.

Рекуррентные нейронные сети (Recurrent Neural Network – RNN) являются специализированными архитектурами в области машинного обучения, предназначенными для работы с данными, представленными в виде последовательности, как, например, временные ряды или текст. Особенностью RNN является наличие обратных связей, которые позволяют сохранять информацию от одного вычислительного шага к другому, имитируя таким образом форму краткосрочной памяти. Это позволяет сети сохранять контекст и учитывать его при обработке каждого нового элемента последовательности.

В отличие от традиционных глубоких нейронных сетей, где информация передаётся строго в одном направлении, RNN способны возвращать информацию обратно по своим слоям. Такие модели, как Long-Short Term Memory, были разработаны для решения проблемы затухания градиента в стандартных RNN, что делает их более эффективными в обучении на длинных последовательностях. LSTM используют специальные структуры, называемые ячейками памяти, что позволяет им не только сохранять информацию на значительный период времени, но и определять, какая информация важна и должна быть сохранена или забыта.

Двунаправленные LSTM (Bi-LSTM) расширяют эту идею ещё дальше, обрабатывая данные как в прямом, так и в обратном направлении, обеспечивая полное понимание контекста во временной последовательности. Такой подход

значительно увеличивает объём контекстуальной информации, доступной для каждой точки данных, что позволяет сети лучше улавливать зависимости и паттерны, возникающие на разных временных интервалах.

Применение RNN и их производных, таких как LSTM и Bi-LSTM, особенно ценно в области биоинформатики для анализа и обработки белковых и геномных последовательностей, где важно учитывать длинные и сложные взаимосвязи между аминокислотами или нуклеотидами. Эти модели особенно подходят для задач, связанных с распознаванием, классификацией и предсказанием биологических функций на основе структурной и последовательной информации.

Комбинация RNN с CNN представляет собой мощный инструмент для глубокого анализа и понимания биологических последовательностей. CNN могут извлекать пространственные признаки из последовательностей, в то время как RNN могут следить за временными или последовательными закономерностями, в результате чего образуются модели, способные обрабатывать и аннотировать сложные биологические данные с высокой степенью точности. Таким образом, объединённая модель может использоваться для эффективного предоставления функциональных аннотаций в системе Gene Ontology.

Полносвязные глубокие сети (Fully Connected Deep Network – FCDN) – это тип нейронных сетей, характеризующихся тем, что каждый нейрон в одном слое связан со всеми нейронами в следующем слое. Это обеспечивает плотную сеть связей, позволяющую модели выявлять и интегрировать информацию из всех входных данных. FCDN нашли широкое применение в различных задачах обработки данных, в том числе для преобразования векторов из баз данных белковых доменов, таких как InterPro, и для предсказания функциональных категорий согласно Gene Ontology.

В контексте методов обучения обучение без учителя отличается от обучения с учителем тем, что оно не зависит от предварительно размеченных данных и способно самостоятельно находить структуру в неорганизованных данных. Эта способность особенно ценна при работе с большими объёмами неразмеченных данных, где отсутствуют явные метки для обучения. Обучение без учителя обеспечивает возможность обнаружения внутренних связей и распределений, которые могут быть неочевидными на первый взгляд.

Модели обучения без учителя применяются для группировки данных по схожим характеристикам (кластеризация), сокращения количества переменных для упрощения моделей (уменьшение размерности), а также для выявления более информативных представлений данных, которые могут быть использованы в последующем обучении с учителем или для интуитивного понимания структуры данных. Эти методы особенно важны для предварительной обработки данных перед использованием в более сложных моделях обучения с учителем и могут значительно улучшить производительность последних путём отбора признаков или снижения шума.

Такие техники как алгоритмы кластеризации, например, *k*-means или иерархическая кластеризация, позволяют исследовать и понимать данные без

предвзятых предположений, идентифицировать подгруппы в данных или уменьшить их сложность для улучшения аналитической точности. Методы уменьшения размерности, включая главные компоненты или t-SNE, облегчают визуализацию многомерных данных, выделяя наиболее значимые векторы изменений и способствуя лучшему пониманию структуры данных.

В добавление к этим методам, методы преобразования данных, такие как автоэнкодеры, используются для создания более эффективных и сжатых представлений данных, которые сохраняют ключевые характеристики исходного набора данных, одновременно устраняя избыточность и шум. Это позволяет создавать более сложные и мощные модели обучения с учителем, которые могут справляться с задачами классификации и предсказания с высоким уровнем точности, что крайне важно в таких областях, как биоинформатика, где требуется высокая точность в прогнозировании биологических функций на основе молекулярных данных.

Автоэнкодеры (Autoencoder – AE) [77] представляют собой специализированный класс нейронных сетей, использующихся для обучения без учителя, и они оказались эффективными в задачах сжатия и восстановления данных, включая распределение функциональных аннотаций Gene Ontology по аминокислотным последовательностям. Принцип работы AE заключается в том, что он учится кодировать входные данные в сжатом виде и затем восстанавливать их до исходного состояния, стремясь минимизировать различия между входом и выходом.

Процесс обучения автоэнкодера включает два основных этапа: кодирование (сжатие) входных данных до представления с меньшим размером и декодирование (восстановление) этого представления обратно до размера исходных данных. Целью такого двойного преобразования является извлечение наиболее существенных характеристик данных, таким образом, чтобы сжатое представление содержало достаточно информации для восстановления исходной информации.

После обучения автоэнкодер может быть использован для предсказания аннотаций GO, используя полученные сжатые представления белковых последовательностей, которые могут содержать ключевые биологические сигналы [78]. Эти сжатые представления также могут служить входными данными для других классификаторов, таких как машины опорных векторов, для проведения окончательной классификации. SVM хорошо подходят для разделения данных на классы, даже если представление признаков имеет высокую степень абстракции.

Кроме того, автоэнкодеры могут работать в сочетании с сверточными нейронными сетями, где сжатые признаки, извлечённые из скрытого слоя AE, подаются в CNN для дальнейшей обработки и классификации. Это сочетание позволяет объединить преимущества обоих подходов: глубокую обработку признаков с помощью CNN и эффективное сжатие данных с помощью AE. В результате, такая интегрированная система способна обеспечить более точное и надёжное предсказание функциональных аннотаций для белковых последовательностей [80].

Ограниченная машина Больцмана (Restricted Boltzmann Machine – RBM) представляет собой двухслойную нейросетевую модель, которая включает в себя слой видимых узлов, кодирующих входные данные, и один скрытый слой, ответственный за извлечение и представление латентных признаков. В контексте аннотации Gene Ontology, видимый слой RBM моделирует распределение входных терминов GO, что позволяет обнаруживать взаимосвязи между различными GO терминами и используемыми последовательностями. В [81] была разработана и описана концепция глубокой RBM (Deep RBM или DRBM), расширяющая стандартную RBM за счёт добавления дополнительных скрытых слоёв. Многоуровневая структура DRBM позволяет модели обучаться последовательно на каждом уровне, что углубляет её способность к обобщению и повышает точность предсказаний функций белков.

Такие глубокие модели, как DRBM, используют жадные алгоритмы пошагового обучения, которые позволяют каждому новому скрытому слою настраиваться на сложные абстракции, выявленные предыдущими слоями. Это ступенчатое обучение способствует построению более комплексной модели, которая может эффективно расшифровывать и предсказывать сложные биологические паттерны и механизмы. Кроме того, применение таких моделей в биоинформатике демонстрирует их потенциал в задачах классификации и функционального аннотирования на молекулярном уровне, позволяя исследователям более глубоко понимать и предсказывать биологические функции на основе генетической информации.

Использование глубоких архитектур, таких как DRBM, обеспечивает значительные преимущества перед традиционными подходами в анализе и интерпретации больших биологических данных. Например, многослойная обработка данных позволяет DRBM фильтровать шум и выделять ключевые сигнатуры, которые могут быть неочевидны на уровне отдельных генов или белков, но которые проявляются на уровне более высоких функциональных групп. Таким образом, DRBM и подобные им глубокие нейросетевые модели открывают новые горизонты для точного предсказания функциональных свойств биологических последовательностей и могут служить мощным инструментом в разработке новых биомедицинских технологий и лекарственных препаратов.

Наконец, генеративно-состязательные сети (Generative Adversarial Network – GAN) представляют собой передовую структуру глубокого обучения, включающую в себя два типа сетей, работающих в своеобразном сотрудничестве: генератор, задача которого заключается в создании искусственных данных, неотличимых от настоящих, и дискриминатор, который стремится отличить эти синтетические данные от истинных. Генератор обучается производить всё более убедительные данные, в то время как дискриминатор становится лучше в их распознавании, что в совокупности способствует улучшению обоих компонентов.

В области автоматического предсказания функций применение GAN может значительно расширить возможности моделей, обученных с учителем, поскольку с помощью GAN создаются высококачественные синтетические данные. Эти данные могут служить дополнительным обучающим материалом

для классификаторов, улучшая их способность к обобщению и увеличивая точность прогнозов. Генераторы в GAN, используя сложные онтологические корреляции, способны синтезировать биологические функции, которые обогащают тренировочные наборы данных и помогают моделям лучше адаптироваться к реальным задачам классификации [82].

Применение GAN в AFP обеспечивает не только генерацию новых, но и расширение существующих наборов данных белковых последовательностей, что может быть особенно ценно в случаях, когда истинные экспериментальные данные ограничены или труднодоступны. Генератор в GAN обучается на основе различных последовательностей и функциональных аннотаций, анализируя и моделируя сложные биологические взаимодействия и свойства, что ведёт к созданию новых искусственных примеров, которые можно использовать для дальнейшего обучения классификаторов.

В то же время, дискриминатор в GAN обучается различать синтетические последовательности от реальных, что повышает уровень понимания характеристик, присущих истинно функциональным биологическим структурам. Эта двойная динамика обучения приводит к созданию сетей, которые могут более точно оценивать и предсказывать биологические функции. Кроме того, интеграция GAN в процессы AFP может способствовать разработке новых подходов к функциональному аннотированию, учитывая онтологические корреляции, что позволяет создавать модели, которые могут улавливать более тонкие и сложные биологические отношения и функции, отраженные в Gene Ontology.

В контексте обработки информации существуют различные методики глубокого обучения для назначения функциональных категорий Gene Ontology неопределённым белкам или генным продуктам. Первый подход фокусируется исключительно на анализе аминокислотных последовательностей и применяется для идентификации потенциальных функций белков, которые не имеют известных гомологов или не связаны с другими уже известными белками. Этот подход особенно полезен в случаях, когда отсутствует информация о родственных последовательностях, тем самым давая возможность обнаружить новые биологические функции, основанные исключительно на первичной структуре белка.

Вторая стратегия глубокого обучения включает в себя использование более обширных данных, которые могут включать не только первичную структуру белка, но также третичную структуру, физико-химические свойства, взаимодействия белков и другую релевантную информацию, собранную из множества научных источников и баз данных. Этот многоаспектный подход может значительно улучшить точность предсказания функций, поскольку он учитывает широкий спектр биологических данных и контекстов, в которых белки действуют. Сочетая эти два подхода, исследователи могут разрабатывать более мощные и точные модели для предсказания функций белков. Такие модели могут быть обучены распознавать сложные шаблоны в больших наборах данных, что делает возможным идентификацию потенциальных функций даже для тех белков, которые ранее не были охарактеризованы.

Одним из первых методов прогнозирования аннотаций GO, основанных на глубинном обучении, является метод, предложенный в [78, с. 533]. Авторами были сравнены два решения для аннотаций, tSVD и нейронная сеть AE. Форма вывода двух методов была одинаковой; однако последний показал лучшие результаты на шести разных наборах данных. Позже было предложено применение глубоких рекуррентных машин Больцмана в качестве развития предыдущих подходов, целью которого было расширить возможности аннотирования генетических продуктов. Эта улучшенная методика была применена для анализа и аннотирования функций генов у четырёх ключевых модельных организмов, которые имеют фундаментальное значение в биологических исследованиях: *Homo sapiens*, *S.cerevisiae*, *Mus musculus* и *Drosophila*. Этот подход не только продемонстрировал эффективность DRBM в обнаружении и классификации биологических функций в широком спектре геномных контекстов, но также обеспечил более глубокое понимание молекулярных механизмов разных видов.

Инструмент ProLanGo [83] первым адаптировал технологии нейронного машинного перевода (Neural Machine Translation – NMT), подобные разработанной Google для языкового перевода, к задачам биологической информатики. В этом процессе аминокислотные последовательности и соответствующие термины Gene Ontology кодируются в уникальные языковые системы – «ProLan» для последовательностей и «GOLan» для GO терминов. Посредством использования модели NMT с трёхслойной RNN-архитектурой ProLanGo генерирует GO аннотации для новых белков, тем самым преобразуя сложные биологические данные в понятные аннотации.

DeepSeq [84] представляет собой ещё один передовой метод в области AFP, который использует мощность свёрточных нейронных сетей для анализа аминокислотных последовательностей и присвоения им соответствующих терминов GO. Несмотря на то, что первоначальные исследования с DeepSeq сосредоточились на прогнозировании пяти наиболее распространённых молекулярных функций белков человека (*Homo sapiens*), метод имеет потенциал для расширения и адаптации к более широкому спектру онтологий и организмов.

DEEPred [24, с. 2] представляет собой инструмент в области биоинформатики, использующий глубокие нейронные сети для решения задач многозадачного обучения. Основываясь на архитектуре DNN с прямыми связями, DEEPred строит отдельные нейронные сети для каждого уровня иерархии терминов Gene Ontology, представленных в формате направленного ациклического графа. Эта модульная структура позволяет тщательно обрабатывать предсказания на различных уровнях сложности, что способствует повышению точности и релевантности итоговых аннотаций. Для представления белковых последовательностей DEEPred использует трёхмерный подход дескрипторов: карту профиля подпоследовательности, состав псевдоаминокислот и функцию объединенной триады. Каждый из этих дескрипторов даёт различные важные характеристики последовательности, причём карта профиля подпоследовательности выявляется как эффективный инструмент для анализа и предсказания функций белка, поскольку она в

значительной мере способна захватить паттерны, присущие белковым последовательностям. Этот многогранный метод анализа усиливает способность DEEPred к точному предсказанию, благодаря чему он может служить надёжным инструментом для определения потенциальной функции белков, особенно в условиях отсутствия полной гомологичной информации.

DeepGOPlus [26, с. 423] был создан для устранения некоторых ограничений, с которыми столкнулась предшествующая модель DeepGO. Среди улучшений, внесённых в DeepGOPlus, значатся повышенная адаптивность к длине белковых последовательностей, более продвинутое использование информации о белок-белковых взаимодействиях, а также расширенный набор меток GO для более детальной аннотации. DeepGOPlus использует сложную конструкцию многослойных свёрточных нейронных сетей, которые эффективно сочетаются с анализом сходства последовательностей для предсказания функциональных аннотаций. TALE представляет собой методологию, которая объединяет элементы кодировщика преобразователя с анализом сходства последовательностей для уточнения предсказаний GO. Этот подход позволяет включить сложные последовательностные шаблоны в модели предсказания, что усиливает их предсказательную способность и точность. PFP-WGAN [85] является одним из последних нововведений, включающих технологии генеративно-состязательных сетей для определения функций белков. Инновация этой системы заключается в параллельной обработке не только экспериментально подтверждённых аннотаций из базы данных Swiss-Prot, но и синтетически сгенерированных данных, созданных сетью-генератором. Эта двойная система позволяет генератору учиться на реальных аннотированных последовательностях, в то время как дискриминатор становится всё более эффективным в отличии подлинных аннотаций от синтезированных. Этот подход может значительно улучшить качество и точность функциональных аннотаций, предоставляемых для новых белковых последовательностей, расширяя при этом спектр возможных функциональных аннотаций, доступных для анализа.

Функциональное назначение аминокислотных последовательностей трехмёрных структур было предложено в [86]. В основном анализе использовалась модель CNN, которая была обучена и протестирована на пяти наборах данных белков человека, каждому из которых было присвоено два термина GO. Однако прогноз не распространялся на дерево DAG для унаследованных терминов GO. DeepGO [87] является сервером для функциональной аннотации белков, использующим свёрточные нейронные сети для анализа аминокислотных последовательностей и взаимодействий белок-белковых взаимодействий. Этот инструмент присваивает GO-метки белкам, используя метод иерархической классификации, который организован в структуре направленного ациклического графа. Вдохновлённый возможностями DeepGO, DeepAdd расширяет функциональность путём включения дополнительных сведений о структуре CNN, позволяя таким образом глубже исследовать векторные представления, полученные из аминокислотных последовательностей, в дополнение к другой информации. Если доступ к

данным PPI отсутствует, DeepAdd компенсирует это, включая в анализ профиль белковой последовательности. На основе аналогичных принципов совмещения данных о первичной структуре белка и PPI была разработана ещё одна модель GONET [88], которая объединяет возможности CNN, рекуррентных нейронных сетей и уровня внимания (Attention layer) для повышения точности предсказаний по аминокислотным последовательностям *Homo sapiens* и *Mus musculus*. Этот подход использует комплексные вычислительные стратегии для выявления скрытых паттернов и зависимостей в биологических данных, что позволяет создавать более точные модели для функционального аннотирования в широком спектре биологических условий. GONET и модели, подобные ей, показывают важность объединения различных источников данных и аналитических подходов в создании мощных инструментов для геномных исследований.

Работая с набором данных, идентичным тому, который был собран в [73, с. 3], была протестирована многозадачная глубокая нейронная сеть, в которой предсказание функции белка рассматривалось как проблема классификации с несколькими метками. Решение состояло из слоёв, общих для всех задач (меток GO), которые сложены параллельно слоям, специфичным для задачи. DeepFunc [89] представляет собой инструмент для предсказания функций белков, который отличается от аналогов, таких как DeepGO, FFPred3 и BLAST, более высокой эффективностью. Информация, касающаяся белковых доменов, семейств и мотивов, сначала извлекается из базы данных InterPro и затем проходит через серию полносвязных слоёв для кодирования. Дополнительно алгоритм Deepwalk используется для генерации топологических особенностей сети белок-белковых взаимодействий. Метод DeepFunc интегрирует оба вида признаков – последовательностные и сетевые – для создания комплексного ввода, который соответствует архитектуре полносвязной глубокой нейронной сети. Архитектура DeepGOA [25, с. 2] является более продвинутой по сравнению с DeepFunc, интегрируя не только данные, полученные от InterPro и PPI, но также извлекая глобальные и локальные семантические признаки аминокислотных последовательностей. Для этого используются методы, основанные на двунаправленных LSTM и свёрточных нейронных сетях. Эти подходы позволяют детально анализировать и интерпретировать широкий спектр биологических данных для точного предсказания функций белков. SDN2GO, используя подобные типы данных, применяет три отдельные подмодели для каждого источника информации, что позволяет добиться комплексного понимания и предсказания функций белков.

В исследовании, посвященном deepNF [79], была разработана архитектура на основе мультимодального глубокого автоэнкодера, цель которой – извлечение сложных скрытых данных из разнообразных белковых взаимодействий в сетях. Этот подход позволяет улавливать глубокие и неочевидные взаимосвязи, которые могут быть не доступны для прямого анализа. В том же исследовательском контексте DeepMNE-CNN показывает улучшенные результаты по сравнению с deepNF, особенно при работе с данными *Homo sapiens*. Это достигается благодаря интеграции свёрточных нейронных сетей, которые превосходят машины опорных векторов в задачах классификации. CNN

эффективно извлекают важные признаки из данных благодаря их способности обрабатывать входные данные в многомерном пространстве. Таким образом, использование CNN вместо SVM в DeepMNE-CNN не только повышает производительность модели в целом, но и обеспечивает более глубокое понимание биологических данных, что ведёт к более точным предсказаниям в классификации белков. FFPred-GAN [90] использует GAN для расширения возможностей традиционных моделей машинного обучения, особенно SVM. Особенности данного инструмента – это биофизическая информация, извлечённая из необработанных аминокислотных последовательностей с помощью FFPred, а генератор использует скрытые переменные для увеличения синтетических образцов.

Несмотря на то, что описанные выше модели обеспечивают относительно хорошие результаты прогнозирования при решении задачи прогнозирования функции белка, всё же существуют некоторые проблемы. С одной стороны, сетевая структура не может эффективно фиксировать долгосрочную зависимость между одной и той же последовательностью белка и не может полностью извлекать информацию о последовательности аминокислот. Долгосрочная зависимость относится к отношениям зависимости на большом расстоянии между каждой аминокислотой в последовательности белка. Установив эту взаимосвязь, можно лучше усвоить общую информацию о последовательности. С другой стороны, трудно эффективно различать достоверную информацию и недостоверную информацию о последовательности белка, а также уловить аминокислотную последовательность, которая оказывает большее влияние на функцию белка. Достоверная информация относится к информации о последовательности белка, которая оказывает большое влияние на функцию белка. Соответственно, недействительная информация относится к информации о последовательности белка, которая оказывает меньшее влияние на функцию белка.

При внимательном изучении и анализе предыдущих работ, посвящённых проблеме аннотации белков, было выявлено, что классические методы прогнозирования функций белков часто затрудняют выявление глубинных (нелинейных) взаимосвязей между белками и функциональными терминами Gene Ontology. По сравнению с традиционными методами машинного обучения, методы глубинного обучения могут учиться на массивных данных о последовательностях белков без разработки признаков. Пока данные аминокислотной последовательности просто обрабатываются, их можно напрямую вводить в нейронную сеть для обучения. Методы глубинного обучения решают проблемы, которые трудно было решить с помощью традиционных алгоритмов машинного обучения в прошлом, таких как высокая размерность, избыточность и высокий уровень шума, вызванные массивными данными о последовательностях белков.

Выводы по разделу

1. Современные биологические базы данных, такие как GenBank, Pfam, ENA, PDA и UniProt на сегодняшний день содержат терабайты данных

последовательностей в различных форматах и с каждым днём их количество только увеличивается. Эти данные представляют огромный интерес для исследователей в сфере биологии и медицины, так как являются ключом к разгадке механизмов болезней и создания лекарств. Объём доступных данных не позволяет проводить качественный анализ путём ручной обработки, поэтому появляется всё больше инструментов для хранения, обработки и анализа генетических и белковых последовательностей. Именно эта потребность является основной причиной создания новых алгоритмов для эффективной обработки и анализа этих данных.

2. Точная идентификация белков и пептидов является одной из основных задач протеомики. На сегодняшний день существует несколько классических поисковых инструментов для идентификации белков в базах данных уже известных белков (Sequest, Mascot, X!Tandem). Однако идентификация осложняется несколькими факторами: один пептид может быть сопоставлен с несколькими белками в базе данных; существует малое количество идентификаций пептидов для каждого белка; частота ложных обнаружений может превышать допустимые значения, что приводит к ненадёжным идентификациям.

3. Другой важной задачей биоинформатики является аннотация белков или предсказание их функций. Современный подход к функциональному аннотированию белков обязательно включает в себя использование проекта Gene Ontology. Существующие алгоритмы для предсказания функций белков показывают неплохие результаты, но во многом могут быть расширены. Наиболее перспективным направлением для этой проблемы являются постановка задачи как классификационной и использование машинного обучения для её решения.

2 РАЗРАБОТКА АЛГОРИТМА ИДЕНТИФИКАЦИИ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ПРИМЕНЕНИЕМ ВОЗМОЖНОСТЕЙ МАШИННОГО ОБУЧЕНИЯ

2.1 Описание процесса получения данных путём масс-спектрометрии и наборов данных для обучения модели

В современной протеомике для идентификации белков и их пептидных фрагментов широко применяется методика масс-спектрометрии. Сначала белки в биологической пробе подвергаются денатурации, а затем ферментативному расщеплению с использованием трипсина, который нацеливается на аминокислотные остатки лизина (K) и аргинина (R), тем самым преобразуя белковые цепи в множество пептидных фрагментов различных длин. Этот процесс позволяет более детально анализировать структуру исходного белка. После стадии расщепления следует очистка смеси пептидов от прочих компонентов, чтобы избежать интерференции в ходе анализа. Далее полученные пептиды подаются в масс-спектрометр, который создаёт спектры масс-заряда для каждого пептида, что позволяет их идентифицировать и определить их количественное содержание в образце. Численные методы идентификации белков, представленные на рисунке 3, опираются на сопоставление масс-спектрометрических данных с теоретическими пептидными картами, сгенерированными из известных белковых последовательностей. Эти методы используют сложные алгоритмы для сравнения экспериментально полученных спектров с базами данных теоретически сгенерированных пептидов [91-93]. Сравнение проводится на основе эвристических функций оценки, включая методы, основанные на скалярном произведении [94], что позволяет оценить степень совпадения между экспериментальными и теоретическими спектрами, количественные методы оценки общего количества ионных пиков [95, 96], что помогает выявить ключевые фрагменты спектров, а также подходы, использующие точное сопоставление ионов, для определения присутствия специфических фрагментов ионов в экспериментальных спектрах. Эти методы подразумевают использование передовых статистических и вычислительных подходов, таких как применение множественных гипотез при сравнении пиков, что обеспечивает более точное и детальное сравнение с учётом специфичности ионного распределения в спектре, а также применение сложных моделей оценки качества сопоставления, что учитывает не только количество и интенсивность ионных пиков, но и их массовые расхождения и пространственное распределение.

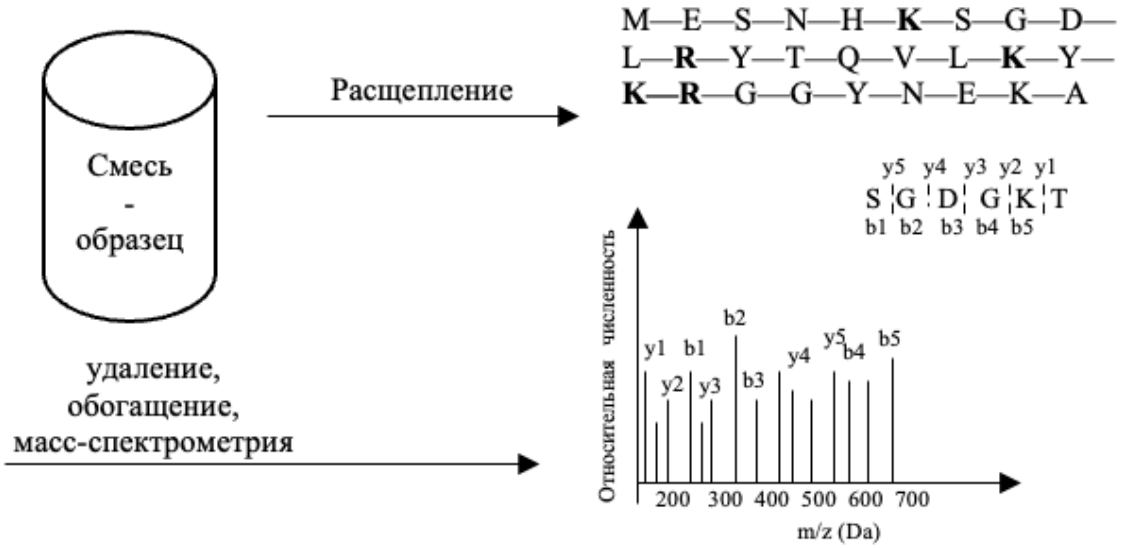


Рисунок 3 – Процесс получения спектров белковой смеси с использованием масс-спектрометрического анализа

Пептиды, построенные из цепочек аминокислотных остатков, представляют собой биополимеры, образующие белки. Каждая аминокислота в цепи, обозначаемая уникальным символом из стандартного набора двадцати различных букв, определяет уникальные свойства пептида. Пептиды находятся в контрасте со спектральными данными, которые являются числовыми рядами, полученными через детальный процесс масс-спектрометрической фрагментации, где каждый пик отображает массу/заряд определённого фрагмента молекулы. Эти два типа данных занимают различные ниши в анализе, и преобразование информации между ними представляет собой сложную задачу аппроксимации и интерпретации. Методы *de novo*, стремящиеся воссоздать первичную структуру пептидов напрямую из масс-спектрометрических данных, сталкиваются с ограничениями из-за шума в данных и частой неполноты фрагментации. Поиск в базе данных, с другой стороны, хоть и обеспечивает более высокую степень точности за счёт сопоставления с уже известными последовательностями, зависит от полноты и точности этих баз, а также от качества экспериментально полученных спектров. Разработка адаптивной системы, которая может синтезировать и анализировать данные, находясь между этими двумя представлениями – численным спектральным и символьным аминокислотным, может значительно улучшить точность идентификации и предсказания белковых функций, что открывает новые горизонты в протеомном анализе.

Наборы данных, необходимые для тренировки и оценки алгоритмов, создаются из спектральных библиотек, собранных и доступных в обширных онлайн-репозиториях, таких как NIST и MassIVE. Эти библиотеки содержат обширные коллекции спектров, собранных и систематизированных на основе различных масс-спектрометрических экспериментов, предоставляя важную основу для исследований и разработок в области протеомики. В ходе

предварительной обработки данных из этих библиотек формируются две основные группы: одна включает в себя сами спектры, а другая – соответствующие пептидные последовательности. Извлечённая таким образом коллекция содержит примерно 4.8 миллиона спектров, сопоставленных с их исходными пептидами, из которых около полумиллиона представляют собой спектры модифицированных пептидов, что добавляет сложности задаче идентификации. Для обеспечения объективности оценки качества алгоритма спектры разделяются на обучающую и тестовую выборки в соотношении 80% к 20%, с тщательным учётом, чтобы исключить повторения или перекрёстное включение пептидов между выборками. Кроме того, для окончательной валидации эффективности обученной нейросети выделяется дополнительный набор данных – отдельная спектральная библиотека, которая не использовалась ни в одном из предыдущих этапов, ни для обучения, ни для тестирования. Данная процедура предварительной подготовки и разделения данных обеспечивает надёжную основу для последующего обучения и проверки алгоритмов, позволяя точно оценить их способность генерировать достоверные результаты в условиях, максимально приближенных к реальной лабораторной практике. Информация о составе и характеристиках спектральных библиотек, выбранных для каждого этапа обучения, тестирования и валидации, предоставляется в виде подробных описаний, включая методы получения данных, условия проведения экспериментов и критерии включения спектров в каждую из библиотек. Формат исходных данных показан на рисунке 4:

Name: AAAAGQTGTVPPGAPGALPLPGMAIVK/2_0_76eV		
MW: 2414.3344		
Comment: Single Pep=SemiTryptic Mods=0 Fullname=A.AAAAGQTGTVPPGAPGALPLPGMAIVK.E Charge=2		
Num peaks: 122		
143.0816	2935.3	b2/0.7ppm
147.1127	2210.5	y1/-0.7ppm
154.4137	1587.1	
155.0816	2825.2	Int/PG/0.6ppm
169.771	1665.2	
176.5942	1535.3	
181.0984	2034.3	
186.0873	2509.4	
198.5815	1458.2	
199.2009	1692.3	
206.924	1799.6	

Рисунок 4 – Формат представления данных масс-спектрометрии

Подробная информация о библиотеках, используемых для обучения, тестирования и проверки представлена ниже.

Обучение/Проверка (Разделение 80/20):

1. human_synthetic_hcd_selected (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:kustersynselected201 70530)

2. cptac2_mouse_hcd_selected (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:clib:mousehcd_selected20141124)
3. chinese_hamster_hcd_selected (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:cho_20180223)
4. human_hcd_tryp_best (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503)
5. massive (https://massive.ucsd.edu/ProteoSAFe/result.jsp?task=daa7c2c21f9a45c08c41e071a3729d67&view=download_filtered_mgf_library&show=true)
6. rat_qtof_consensus_final_true_lib (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:rat_qtof)
7. human_hcd_tryp_good (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503)
8. human_hcd_semitryp (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:humanhcd20160503)
9. human_hcd_labelfree_phospho_selected_passed (https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:phoshopept_labelfree_20190214
10. proteome tools (http://www.proteometools.org/index.php?id=53)

Обучающий набор данных содержит выборку из девяти посттрансляционных модификаций белков, каждая из которых играет уникальную роль в биологических функциях и структурных изменениях пептидов. Для эффективной тренировки нейросетевой модели эти модификации, такие как карбамидометилирование, фосфорилирование, окисление метионина и ацетилирование на *N*-конце, вводятся в процесс обучения как отдельные параметры, каждому из которых соответствует определённый символ в пептидной последовательности. Эти символы и связанные с ними модификации детально перечислены в таблице 3, которая служит своего рода шифровальным ключом для преобразования модификаций в коды, распознаваемые машиной. Сеть обучается распознавать и интерпретировать каждую из этих специфических модификаций, что позволяет ей адаптироваться к различным вариациям пептидных цепей, что важно для точного определения их структуры и функции. Состав обучающего набора данных включает в себя пептидные последовательности, полученные из таких источников, как спектральная библиотека NIST и репозиторий MassIVE, отвечающих за сбор и систематизацию спектрометрических данных масс-спектрометрии. Дополнительные сведения об этих источниках, а также об использованных методах обработки данных представлены в таблице 4, которая предоставляет более глубокое понимание об основе для обучения модели. Эти данные включают в себя весь комплекс посттрансляционных модификаций, таких как карбамидометилирование, что связано с защитой цистеиновых остатков, фосфорилирование, являющееся ключевым механизмом регуляции белковой активности, окисление метионина, указывающее на окислительный стресс, и *N*-концевое ацетилирование, влияющее на стабильность белка и его

взаимодействие с другими молекулами. Полученные и систематизированные таким образом данные обучающего набора представляют собой фундамент для построения нейросетевой модели, способной с высокой точностью идентифицировать и классифицировать пептидные последовательности и их модификации.

Таблица 3 – Модификации и значения символов, используемые в обучающих данных

Модификация	**Символ**
Phospho	p
Oxidation	o
Deamidation	h
Carbamidomethyl	c
Acetyl	a
Ammonia-loss	r
Carbamyl	y
Dehydrated	d
Delta:H(2)C(2)	t

Таблица 4 – Характеристики используемого обучающего набора данных

Параметры	**Значения**
Количество обучающих образцов	4 800 000
2	2 600 000
3	1 600 000
4	400 000
Другие	1 200 000
Немодифицированные образцы	4 300 000
Модифицированные образцы	500 000
Максимальный	8
Количество видов	7

В ходе подготовки данных к тренировке нейросети, спектры масс/заряда (*m/z*) преобразуются в высокоразмерные векторные представления фиксированной длины, каждое из которых охватывает 80 000 метрических единиц. Эти векторы представляют собой дискретизированное отображение спектра с шагом 0,1 Да, обеспечивая детальное и точное представление спектра вплоть до 8000 Да. В этом процессе значения интенсивности каждого отрезка *m/z* подвергаются стандартизации, так что итоговые данные имеют нулевое среднее и стандартное отклонение, равное единице. Подход к нормализации и дискретизации данных спектров обеспечивает унификацию и согласованность данных перед их вводом в сеть, позволяя модели более эффективно обучаться и извлекать закономерности из предоставленного набора данных. Нормализация интенсивности к стандартным значениям позволяет уменьшить влияние случайных колебаний и упростить выявление ключевых паттернов в данных. Для обеспечения единообразия входных данных строки пептидных

последовательностей дополняются нулевыми символами до длины 64 символов, что соответствует стандартной практике подготовки текстовых данных для обучения нейросетей. Это также обеспечивает согласованность размеров данных, важную для параллельной обработки и оптимизации ресурсов. При организации тренировочного процесса данные разделяются на партии по 1024 образца в каждой, что облегчает их обработку и ускоряет процесс обучения.

Модель реализована на языке Python, процесс обучения модели осуществляется на облачной платформе Colab (https://colab.research.google.com/) с помощью открытой библиотеки машинного обучения PyTorch, которая обеспечивает возможность использования последних достижений в области искусственного интеллекта и глубокого обучения, а также поддерживает вычисления на GPU, что значительно ускоряет процесс обучения моделей.

2.2 Реализация алгоритма идентификации пептидов и обучение реализованной модели

В этом исследовании произведено расширение и доработка общедоступной сети подобия SpeCollate [97] для изучения функции подобия для совпадений пептидного спектра. Сети подобия, широко применяемые в различных доменах компьютерного зрения, таких как поиск изображений по текстовому запросу [98] и идентификация лиц [99] и других задач [100, 101], начинают находить своё применение и в сфере протеомики, хотя их использование здесь было до сих пор сравнительно ограниченным. В протеомике, сети подобия используются для поиска по спектральным библиотекам [102] и для группировки спектров по схожести, позволяя классифицировать пептиды по их масс-спектральным характеристикам [103]. Цель такого применения заключается в изучении и создании вложений фиксированной размерности для экспериментальных спектров и переменной длины для пептидных цепочек, таким образом чтобы соответствующие друг другу спектр и пептид были проецированы в близкое пространство, если рассматривать евклидово расстояние (*L2*) как меру их близости. Евклидово расстояние *L2* было выбрано как функция подобия, поскольку исследования показали его эффективность в задачах ранжирования по сходству и его превосходство над другими метриками подобия. Это свидетельствует о том, что векторное представление фрагментов спектра и пептидных цепочек в общем вложенном пространстве может обеспечить более точное сопоставление, что имеет важное значение для идентификации белков и последующего анализа их функций. Такие сети подобия обладают потенциалом для создания новых возможностей в протеомике, обеспечивая высокоточное сопоставление экспериментальных спектров с пептидными цепочками.

Сетевая архитектура SC_MS_Peptide_Ident, представленная для данного исследования, интегрирует две специализированные подсети для обработки данных масс-спектрометрии: спектральную подсеть SSN, включающую два полносвязных слоя для анализа масс-спектров, и пептидную подсеть PSN с одним двунаправленным слоем LSTM и последующими полносвязными слоями для обработки пептидных последовательностей. Обе подсети совместно

обучаются на данных, состоящих из разреженных спектральных векторов и закодированных строк пептидов, где целью является минимизация функции потерь на основе секстиплетов, формируемых по ходу обучения. В частности, функция потерь вычисляется с использованием секстиплетной выборки, включающей якорный спектр Q, соответствующий ему положительный пептид P, а также две отрицательные пары Q_N, P_{N_Q} и Q_N, P_{N_P} соответственно для каждого из них, отобранные с использованием метода онлайн-отбора трудноотделимых отрицательных примеров. Этот подход направлен на повышение точности обучения, поскольку отрицательные спектры и пептиды, которые больше всего похожи на положительные образцы Q и P в каждой партии, выбираются в ходе прямого прохода, тем самым обеспечивая более высокую дискриминативность и ускорение сходимости обучения. Всё это позволяет модели лучше различать положительные и отрицательные примеры, тем самым улучшая качество обучения и повышая вероятность более точного сопоставления спектров и пептидов. Такая сложная схема обучения предназначена для извлечения более глубоких взаимосвязей между спектрами и пептидными последовательностями, что является ключевым для успешного прогнозирования функций белков на основе их масс-спектральных данных.

Полная архитектура сети показана на рисунке 5. Ввод спектров Q в спектральную подсеть осуществляется путём их представления в формате нормализованных разреженных векторов, что способствует увеличению эффективности обработки данных и сокращению необходимой вычислительной мощности. Пептиды (P, N) подаются в пептидную подсеть последовательно в прямом и обратном направлении, что позволяет LSTM-блокам анализировать последовательность как целое, улавливая не только последовательные, но и обратные зависимости между аминокислотами, тем самым обеспечивая более полное и точное предсказание функций белков.

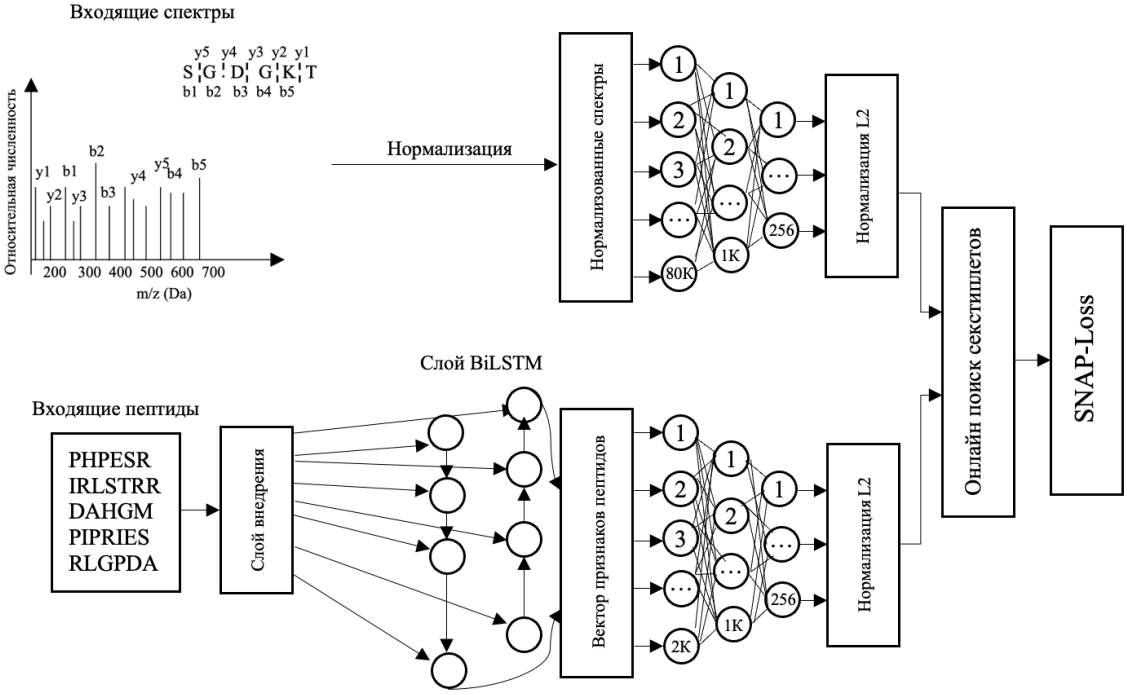


Рисунок 5 – Архитектура сети глубокого подобия

Спектральная подсеть SC_MS_Peptide_Ident служит для глубокого анализа масс-спектров и их преобразования в компактное евклидово векторное пространство размерностью $\mathbb{R}^{256}$. SSN включает в себя два полносвязных слоя, где первый слой принимает 80 000 входных параметров, соответствующих обработанным спектрам, и сокращает их до 1024 признаков, а второй слой дополнительно сжимает информацию до 256 признаков, представляющих собой высокоуровневые абстракции изначального входного спектра. Каждый спектр входит в сеть в виде разреженного вектора, представляющего интенсивности пиков, нормализованных для обеспечения стабильности и согласованности данных, с шагом массового спектра в 0,1 Да. Для активации нейронов в скрытых слоях используется функция ReLU, которая способствует более эффективному обучению сети, позволяя модели быстрее и эффективнее улавливать нелинейные зависимости между данными. Дополнительно применяется dropout с вероятностью отключения нейронов в 30%, что снижает риск переобучения сети путём искусственного ограничения количества активных нейронных связей во время обучения. Это позволяет сети обучаться на более обширном и разнообразном наборе признаков, что в итоге повышает её обобщающую способность при классификации новых спектров. Завершающий слой SSN выполняет нормализацию *L2,* что означает проецирование векторов признаков на поверхность единичной гиперсферы, что упрощает изучение расстояний между точками в пространстве признаков и обеспечивает более стабильное распределение признаков для дальнейшего анализа и классификации.

Подсеть обработки пептидов в модели служит для изучения и представления пептидных последовательностей в том же векторном пространстве, что и спектры, обеспечивая возможность непосредственного

сравнения и соотнесения пептидов с их масс-спектральными профилями. Архитектура PSN содержит двунаправленный слой LSTM (BiLSTM), который оснащен возможностью захватывать взаимосвязи в обоих направлениях вдоль аминокислотной последовательности. Это позволяет модели эффективно интегрировать контекстную и позиционную информацию для каждой аминокислоты, обеспечивая тем самым более полное понимание структуры и функций белков. Перед обработкой в BiLSTM, аминокислотные последовательности проходят через слой встраивания, который преобразует каждый символ аминокислоты в 256-мерный вектор с плавающей точкой, тем самым предоставляя компактное представление каждой аминокислоты. Два последующих полносвязных слоя уплотняют информацию, извлечённую из BiLSTM, сначала до промежуточного представления, а затем до встраивания на единичной гиперсфере, аналогичной спектральной подсети.

Таким образом, в архитектуре модели двунаправленная LSTM применяется для эффективного обучения на основе секвенированных пептидов, включающих в себя стандартные аминокислоты и модификации, обозначенные в специальном словаре размером 30 символов. Словарь составляют 20 стандартных аминокислот, 9 модификаций, каждая из которых имеет уникальное обозначение, и пробел, используемый как заполнитель для выравнивания последовательностей. Скрытые слои BiLSTM, имеющие размерность 1024, осуществляют обработку последовательностей в двух направлениях, что позволяет захватывать контекст аминокислот как до, так и после каждой позиции. Объединённые данные с прямого и обратного проходов формируют вектор длиной 2048, что обеспечивает полное представление динамических особенностей пептидной последовательности. Дальнейшая обработка осуществляется двумя полносвязными слоями, которые сжимают информацию до компактных векторных представлений размерами 1024 и затем 256. Эти слои используют функцию активации ReLU, способствующую нелинейному преобразованию данных, что улучшает способность сети к обобщению и распознаванию сложных паттернов в данных. Механизм dropout с вероятностью 0,3 применяется после каждого из этих слоёв для предотвращения переобучения, обеспечивая таким образом регуляризацию и улучшение обобщающей способности модели. Эти меры предосторожности позволяют сети более эффективно извлекать существенные признаки из пептидных последовательностей и предотвращают переобучение на особенностях обучающего набора данных, которые не являются обобщающими для неизвестных данных.

При реализации модели были использованы следующие параметры: использованы один слой BiLSTM и два полностью связанных слоя, в качестве функции логистических потерь SNAP использована модифицированная функция Triplet-loss.

Процесс обучения сети подобия для идентификации пептидов из масс-спектров представляет собой сложную последовательность шагов, начиная с предварительной подготовки данных и заканчивая реализацией алгоритмов для оптимизации процесса обучения. В ходе этого процесса изначально

используется подход, при котором экспериментальные спектры и соответствующие им пептиды группируются в пакеты по 1024 точки данных и подаются на вход двум подсетям: спектральной и пептидной соответственно. В этот момент отрицательные примеры ещё не сформированы, что означает, что исходный набор данных состоит только из положительных пар спектр-пептид. Следующим этапом в процессе обучения является развитие метода для динамической генерации негативных примеров, что предполагает использование методики онлайн-майнинга секстиплетов. В этом контексте под "секстиплетом" понимается совокупность положительных и отрицательных пар, связанных определённым образом. Для каждой положительной пары спектра (q_i из набора Q) и пептида (p_i из набора P), где Q обозначает множество векторных вложений спектров, а P – множество векторных вложений пептидов, активно подбираются четыре наиболее подходящих негативных примера. Этот процесс отбора базируется на принципе поиска векторных соседей в многомерном пространстве, где для каждой положительной пары (q_i, p_i) выявляются те отрицательные спектры (q_j) и пептиды (p_k), которые являются ближайшими к q_i и p_i соответственно, но не образуют корректные пары друг с другом. Также проводится поиск отрицательных примеров (q_l, p_m) для каждого пептида p_i, где q_l– это спектр, ближайший к p_i, а p_m – пептид, ближайший к p_i, опять же с условием отсутствия соответствия. Целью такого подхода является усиление обучающего эффекта за счёт введения отрицательных примеров, которые находятся в непосредственной близости к положительным в векторном пространстве, при этом не являясь соответствующими друг другу. Это позволяет системе более эффективно различать правильные соответствия от неправильных, повышая качество обучения. Такой механизм генерации и отбора отрицательных примеров помогает модели более точно настраивать векторные вложения для дальнейшего улучшения способности распознавать и отличать положительные пары от отрицательных. Это, в свою очередь, повышает общую эффективность процесса идентификации спектр-пептидных совпадений.

Этот процесс обеспечивает, что сеть не только учится различать правильные спектр-пептидные пары от неверных, но и способствует улучшению способности модели различать близкие, но не соответствующие друг другу примеры, тем самым повышая точность идентификации. После создания секстиплетов, содержащих якорный спектр, положительный пептид и четыре отрицательных примера, модель проходит через фазу обратного распространения ошибки, в ходе которой оптимизируются веса сети для минимизации ошибок классификации. Величина потерь, используемая для оценки качества обучения, вычисляется на основе расстояния между положительными и отрицательными примерами, что позволяет модели более точно оценивать близость между спектрами и пептидами.

Математическое обоснование негативного онлайн-майнинга для генерации секстиплетов может быть представлено следующим образом: входящая партия образцов разделена на две группы: Q и P. Первая группа подвергается обработке с использованием спектральной подсети модели, в то

время как вторая проходит через пептидную подсеть. В результате получаем встроенные спектры Q и пептиды P, которые математически представлены функциями $Q = f_{SSN}(Q)$ и $P = f_{PSN}(P)$, соответственно, где каждый элемент из Q и P является вектором в 256-мерном пространстве, обозначенном $\mathbb{R}^{256}$.

Для повышения эффективности процесса исследования вычисляются отрицательные примеры для каждой положительной пары, состоящей из элементов (q_i из Q и p_i из P), через создание трёх отдельных матриц расстояний: $D_{Q\times Q}$, $D_{Q\times P}$ и $D_{P\times P}$. Эти матрицы включают в себя значения *L2*-расстояний для всех возможных пар комбинаций элементов в выборках, возведённые в квадрат. Таким образом, $D_{Q\times Q}$ отражает квадрат евклидова расстояния между всеми возможными парами спектров $\|q_i - q_j\|^2$, $D_{Q\times P}$ показывает квадрат евклидова расстояния между спектрами и пептидами $\|q_i - p_j\|^2$, а $D_{P\times P}$ демонстрирует квадрат евклидова расстояния между всеми парами пептидов $\|q_i - p_j\|^2$, где i и j являются индексами элементов и могут принимать значения от 1 до значения количества элементов во входящей партии e. Матрицы расстояний, о которых идет речь, представляют собой квадратные симметричные матрицы с размерами $e \times e$, где e указывает на количество элементов в каждой группе. В таких матрицах особое значение имеет расположение элементов: на диагонали матрицы $D_{Q\times P}$ находятся числа, соответствующие квадрату евклидова расстояния между парами, которые считаются совпадающими или положительными парами. Эти значения отражают степень сходства или различия между соответствующими элементами двух групп. В то же время, на диагоналях матриц $D_{Q\times P}$ и $D_{P\times P}$, которые сравнивают элементы внутри одной группы спектров (Q) или пептидов (P) соответственно, располагаются нули. Эти нули символизируют полное совпадение элементов, поскольку расстояние от элемента до самого себя всегда равно нулю, подчёркивая их идентичность в рамках данной группы. Дополнительно, для более глубокого анализа взаимных расстояний внутри группы Q можно воспользоваться понятием матрицы Грамиана, обозначаемой как G_Q. Матрица Грамиана – это матрица, элементы которой представляют собой скалярные произведения векторов, и она используется для анализа внутренней структуры и взаимосвязей внутри группы. Эта матрица представляет собой квадратную таблицу внутренних произведений элементов Q, где каждый элемент матрицы определяется как (5). Кроме того, диагональ матрицы Грамиана, обозначаемую как g_Q, можно вычислить как (6):

$$G_Q = Gramian(Q) = \left[\langle q_i, q_j \rangle\right] \tag{5}$$

$$g_Q = diag(G_Q) \tag{6}$$

Определив матрицу расстояний для группы Q, обозначаемую $D_{Q\times Q}$, можно воспользоваться следующей формулой (7):

$$D_{Q\times Q} = g_Q 1^T - 2G_Q + 1g_Q^T \quad (7)$$

где g_Q представляет собой вектор, состоящий из суммы квадратов каждого элемента из Q, а 1 представляет собой вектор-столбец, содержащий исключительно единицы, размерность которого соответствует длине вектора g_Q, то есть e. Данное уравнение отражает методику вычисления попарных квадратов евклидовых расстояний между всеми спектрами в группе Q, обеспечивая таким образом комплексный подход к измерению вариабельности внутри данной выборки.

Чтобы получить матрицу расстояний между спектрами Q и пептидами P, можно использовать аналогичные операции, которые могут быть представлены как (8), (9) и (10):

$$G_P = Gramian(P) = \left[\langle p_i, p_j \rangle\right] \quad (8)$$

$$g_P = diag(G_P) \quad (9)$$

$$D_{Q\times P} = g_Q 1^T - 2Q^T P + 1g_P^T. \quad (10)$$

Для (8) учитываем, что матрица Грамиана G_Q служит основой для расчёта внутреннего взаимодействия между спектрами. Она формируется путём умножения транспонированного вектора спектров на исходный вектор спектров. Далее, для (9) рассматриваем процесс вычитания удвоенной матрицы Грамиана из произведения вектора g_Q на транспонированный вектор единиц. Наконец, (10) дополняет расчёт добавлением произведения вектора единиц на транспонированный вектор g_Q, что позволяет получить полную матрицу расстояний для группы Q.

Теперь, когда матрицы расстояний подготовлены, следующий этап работы включает поиск и определение негативных примеров. Это достигается посредством функции *min*, применяемой к каждой из матриц. Обозначим элементы матриц $D_{Q\times Q}$, $D_{Q\times P}$ и $D_{P\times P}$ как qq_{ir}, qp_{ir} и pp_{ir} соответственно, где i и r указывают на индексы строк и столбцов соответствующих матриц. Следуя этому подходу, индексы для негативных примеров в секстиплете S могут быть определены по следующим четырём уравнениям (11), (12), (13), (14):

$$j_i = \underset{r, r \neq i}{\arg\min} q_{ir}, \; i = 1, \ldots, e \quad (11)$$

$$k_i = \underset{r, r \neq i}{\arg\min} qp_{ri}, \; i = 1, \ldots, e \quad (12)$$

$$l_i = \underset{r, r \neq i}{\arg\min} qp_{ir}, \; i = 1, \ldots, e \quad (13)$$

$$m_i = \underset{r, r \neq i}{\arg\min}\, p_{ir},\ i = 1, \ldots, e. \quad (14)$$

Эти индексы являются указателями на соответствующие спектры и пептиды с наименьшим расстоянием до каждой точки, исключая саму точку, что важно для определения негативных примеров, используемых при вычислении функции потерь. После того как секстиплеты сформированы, вычисление функции потерь осуществляется с помощью специализированной функции потерь SNAP. Обновление градиентов применяется к обеим нейросетям, что позволяет обратное распространение корректировок для улучшения модели. При онлайн-майнинге секстиплетов, который демонстрируется на рисунке 6, на каждом этапе итерации из пакета выбираются четыре ближайших негативных значения к текущим значениям q или p. В процессе обновления градиента эти негативные значения отдаляются, что заставляет сеть выбирать новые, более сложные негативные примеры на следующей итерации. Этот процесс постоянного обновления негативного набора способствует тому, чтобы сеть училась на наиболее трудных примерах, что способствует оптимизации процесса обучения.

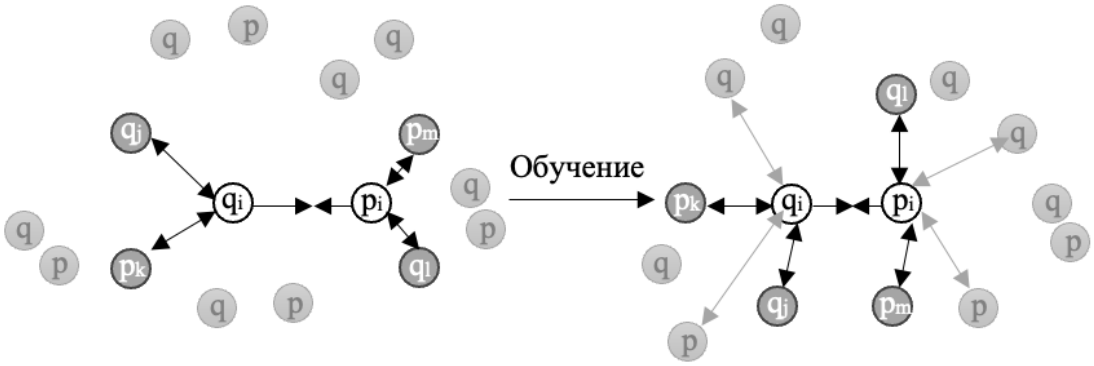


Рисунок 6 – Поиск негативных примеров для функции потерь

Основная задача, которую ставит перед собой процесс обучения модели, заключается в уменьшении квадрата евклидова расстояния между парой, состоящей из спектра и согласованного с ним положительного пептида, при одновременном увеличении этого расстояния для пар спектр-пептид, которые являются несоответствующими или отрицательными. Таким образом, модель стремится к усилению различий между этими двумя категориями, чётко разграничивая классы положительных и отрицательных пар. Методика, применяемая для достижения указанной цели, схожа с концепцией функции Triplet-Loss [104], которая фокусируется на тройках объектов. В рамках данной концепции каждая тройка состоит из трёх элементов: A (*Anchor*) – опорный элемент, P (*Positive*) – положительный элемент, который должен соответствовать опорному, и N (*Negative*) – отрицательный элемент, который не соответствует

опорному. Функция потерь триплетов затем стимулирует модель к тому, чтобы расстояние между A и P было как можно меньше, в то время как расстояние между A и N должно быть как можно больше. Это пространственное разделение помогает улучшить качество классификации, обеспечивая более точное разграничение между согласованными и несогласованными спектральными и пептидными парами в задачах масс-спектрометрической идентификации.

Такой подход обеспечивает более надёжное распознавание и ассоциацию спектров с их истинными пептидными последовательностями, повышая вероятность корректного определения протеомных составляющих образца. Функция Triplet-Loss оценивает каждую тройку на предмет соответствия расстояний между A и P и между A и N заданным порогам, что в итоге ведёт к формированию пространства признаков, где аналогичные объекты сгруппированы вместе, а различные – разделены настолько эффективно, насколько это возможно. Это обучение на контрастах между схожими и различными парами позволяет модели выработать устойчивые паттерны распознавания, что незаменимо в сложных задачах анализа биологических данных.

Функция старается уменьшить разность квадрата евклидова расстояния между привязкой и положительным примером $\|A - P\|^2$ в то время как разность между привязкой и отрицательным примером $\|A - N\|^2$ стремится быть максимально большой. Эта функция потерь включает запас между положительными и отрицательными парами, чтобы обеспечить достаточное пространственное разделение между ними. Этот запас добавляется к расстоянию между привязкой и положительным примером. Таким образом, цель обучения модели – минимизировать значение функции потерь, которое может быть представлено следующим уравнением (15):

$$Loss = -\frac{1}{e}\sum_{i=0}^{e} \max\left(\|A - P\|^2 - \|A - N\|^2 + \text{margin}, 0\right). \quad (15)$$

Методы, основанные на функции потерь Triplet-Loss, оказались чрезвычайно эффективными при работе с данными одной модальности, такими как изображения в системе распознавания лиц, подобной FaceNet [105]. FaceNet – это система, которая использует глубокое обучение для встраивания изображений лиц в евклидово пространство, где расстояния напрямую соответствуют степени схожести лиц. Применяя Triplet-Loss, FaceNet эффективно обучается отличать лица разных людей, приближая изображения одного и того же лица и отдаляя изображения разных лиц в векторном пространстве, что идеально подходит для задач проверки и идентификации.

Функция SNAP-потерь представляет собой модификацию классической Triplet-Loss, адаптированную для работы с мультимодальными данными. В случае числовых спектров и последовательностей пептидов, SNAP-потери предполагают учёт всех потенциальных негативных образцов для заданной положительной пары. Это позволяет улучшить дифференциацию между

положительными и отрицательными примерами в пространстве признаков, таким образом повышая качество модели.

В данной схеме, q_i и p_i представляют собой положительные спектр и пептид соответственно. Отрицательные примеры q_j, p_k, q_l и p_mпредставляют спектры и пептиды, которые не соответствуют положительной паре и выбираются так, чтобы максимизировать общий проигрыш, то есть увеличить разность между положительной и отрицательной парой в пространстве признаков. Усреднение проигрышей по всем отрицательным примерам позволяет обучить модель различать положительные примеры от широкого спектра отрицательных, что может быть важно для задач, где существует большая вариативность в данных.

Функция SNAP-потерь, адаптирующая концепцию Triplet-Loss для мультимодальных данных, здесь применяется к числовым спектрам и последовательностям пептидов. Для расчёта этой функции потерь важно определить и предварительно вычислить несколько ключевых переменных, базирующихся на расстояниях между положительными и отрицательными примерами. Представленные переменные следующие – квадрат евклидова расстояния между положительной парой спектра и пептида, квадрат расстояния между положительным спектром и отрицательным спектром для q_i, квадрат расстояния между положительным спектром и отрицательным пептидом для q_i, квадрат расстояния между положительным пептидом и отрицательным спектром для p_i, квадрат расстояния между положительным пептидом и отрицательным пептидом для p_i (16):

$$\begin{aligned} d_i &= \|q_i - p_i\|^2 \\ d_{n1} &= \left\|q_i - q_j\right\|^2 \\ d_{n2} &= \|q_i - p_k\|^2 \\ d_{n3} &= \|p_i - q_l\|^2 \\ d_{n4} &= \|p_i - p_m\|^2 \end{aligned} \tag{16}$$

Тогда общая функция SNAP-потерь рассчитываются для партии размера e следующим образом (17):

$$Loss = \frac{1}{4e}\sum_{i=1}^{e}\sum_{r=1}^{4}\max\left(d_i - d_{nr} + margin, 0\right) \tag{17}$$

После обучения нейросети она может использоваться для встраивания новых данных в пространство признаков, где объекты, которые рассматриваются как "похожие" на основании изученных нейросетью характеристик, будут группироваться вместе. Для тестового набора данных пептиды и спектры преобразуются в векторы во встроенном подпространстве, что позволяет

использовать методы поиска ближайшего соседа для определения сходства. Индексация встроенных векторов пептидов – это эффективный способ ускорить процесс поиска, особенно когда количество спектров и пептидов велико. Векторы признаков для пептидов могут быть заранее вычислены и сохранены в индексированной структуре данных, такой как хэш-таблица, что позволяет быстро находить наиболее похожие пептиды для заданного спектра. Также рационально выполнить предварительное вычисление и сохранение векторов признаков для экспериментальных спектров. Это избавляет от необходимости повторно кодировать каждый спектр при многократном сопоставлении с пептидами, тем самым существенно ускоряя процесс поиска. Эти оптимизации особенно важны в вычислительных биологии и биоинформатике, где объёмы данных часто очень велики и требуют высокой производительности вычислительных систем.

Использование графического процессора для вычисления мер расстояния, таких как *L2* или евклидово расстояние, значительно ускоряет обработку больших наборов данных, что является типичным для биоинформатики и протеомики. GPU хорошо подходят для параллельных вычислений, поскольку они могут обрабатывать множество операций над векторами и матрицами одновременно. Матрица замаскированных расстояний, которая учитывает только пептиды в определённом диапазоне *m/z* предшественника, может быть эффективно вычислена на GPU. Маскирование позволяет исключить из расчёта все ненужные пары, что дополнительно ускоряет процесс. То есть, если масса предшественника для определённого пептида известна, можно рассчитывать расстояния только для тех спектров, которые попадают в этот диапазон *m/z*, игнорируя остальные. Инвертированное значение расстояния *L2*, то есть *L2 – (1/L2)*, может быть использовано как мера совпадения: чем меньше расстояние, тем выше обратное значение и тем более похожими считаются спектр и пептид. Это даёт количественную меру схожести, которая может быть использована для оценки качества совпадения между спектром и потенциальными пептидами в базе данных.

Используя матрицу расстояний $D_{A\times B}$ и матрицу маски M, можно эффективно вычислять расстояния *L2* только между релевантными пептидами и спектрами, сокращая вычислительную нагрузку и количество данных для анализа. Матрица $D_{A\times B}$ заполняется на основе вычисления евклидовых расстояний между спектрами и пептидами, где каждая строка представляет отдельный спектр, а каждый столбец – отдельный пептид. Элементы этой матрицы определяются через вычисление суммы квадратов элементов для каждого спектра и пептида, вычитание двойного произведения матриц $A_{1024\times 256}$ и $B_{<16384\times 256}$, и добавление квадратов элементов для пептидов (18):

$$D_{A\times B} = g_A 1^T - 21A^T B + 1g_B^T \tag{18}$$

где g_A и g_B – это вектора, содержащие суммы квадратов элементов каждого вектора из спектров и пептидов соответственно. Матрица маски M используется для замаскирования тех расстояний, которые не попадают в заданное окно

массы/заряда предшественника, исключая их из дальнейшего анализа. Произведение Адамара, или поэлементное умножение $D_{A\times B}$ на M, позволяет избежать лишних вычислений для нерелевантных пар спектр-пептид. В результате получается фильтрованная матрица расстояний, которая содержит только значимые данные для анализа. Для каждого спектра затем выбираются пептиды с наименьшим расстоянием, что указывает на наивысшую степень совпадения. Для каждого спектра сохраняются 5 пептидов с наивысшей оценкой (минимальное расстояние), а остальные отбрасываются, что даёт результирующую матрицу оценок размером 1024×5, которая сохраняется для последующего анализа позже. Такой подход упрощает поиск наиболее подходящих пептидов для данных спектров и позволяет эффективно использовать ограниченные ресурсы памяти и вычислительную мощность GPU, что критически важно при работе с большими наборами данных.

2.3 Результаты обучения и оценка разработанной модели

Для обучения модели SC_MS_Peptide_Ident в разных экспериментах было запущено от 20 до 100 эпох обучения с размерностью от 64 до 256 с целью достижения высокой точности классификации, что подтвердило её способность корректно идентифицировать правильные пептиды по спектрам. Точность, или показатель Accuracy, можно определить как долю случаев, когда предсказанный ближайший пептид совпадает с истинным пептидом, связанным с якорным спектром. Функция $tp(q, E)$ является индикатором успешности классификации для каждого спектра q в партии E, возвращая 1, если ближайший пептид p к q является истинным пептидом p_q, и 0 в противном случае (19):

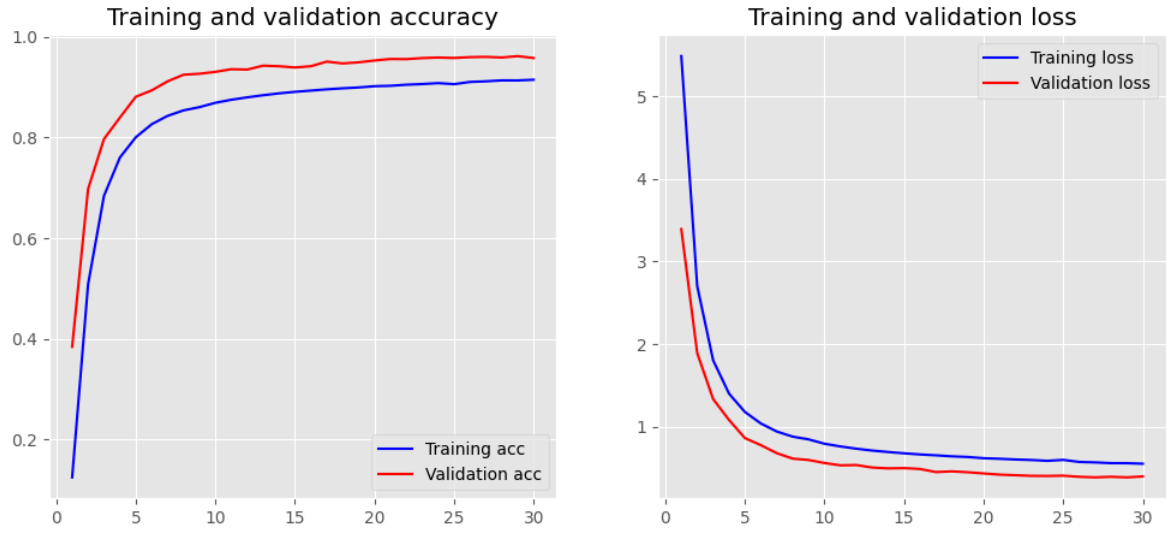
$$tp(q,E) = \begin{cases} 1 \arg\ min_{p\in E} \|q-p\|^2 = p_q \\ 0 \qquad\qquad\qquad otherwise \end{cases} \tag{19}$$

$$Accuracy = \frac{\sum_{q\in E} tp(q,E)}{|E|}$$

где E обозначает количество спектров в партии. Подход использует концепцию ближайшего соседа в векторном пространстве встроенных представлений, где расстояние между точками (в данном случае между спектром и пептидами) измеряется в евклидовом смысле $\|q-p\|^2$. Если минимальное расстояние до любого пептида в партии равно расстоянию до истинного пептида p_q, классификация считается успешной. Эта мера позволяет оценить, насколько хорошо обученная сеть способна распознавать правильные пептиды в новых данных, что является ключевым показателем её эффективности при практическом применении.

На представленном рисунке 7 демонстрируется детализированная визуализация процесса обучения нейронной сети, включая наглядное

представление изменения точности на протяжении установленного количества эпох обучения. График демонстрирует, как точность на обучающей выборке (Training Accuracy) постепенно увеличивается, отражая способность модели все более точно классифицировать обучающие данные. Параллельно с этим, представлены данные по точности на проверочной выборке (Test Accuracy), которая также увеличивается, что свидетельствует об улучшении способности модели к обобщению на данных, которые не использовались в процессе обучения. Кроме того, на графике отображается динамика функции потерь (Loss Function), ключевого компонента процесса обучения, который количественно отражает разницу между предсказаниями модели и истинными значениями. Сокращение значения функции потерь с течением времени указывает на то, что модель становится всё более эффективной в предсказаниях, с каждой эпохой снижая ошибку. Этот важный показатель обучения представлен в динамике, что позволяет наблюдателям оценить прогресс модели и её текущее состояние, а также помогает идентифицировать любые признаки переобучения или недообучения на основе анализа траектории значений потерь и точности во времени.



a b

Рисунок 7 – Результаты обучения сети: *a* – Прогресс обучения для данных обучения/тестирования, *b* – значение потерь

Для оценки эффективности исследуемой модели было проведено тестирование на ряде экспериментальных данных. Комплексное тестирование включало анализ набора данных, предоставленного Национальным институтом стандартов и технологии, а также трёх дополнительных наборов данных, идентифицированных как PXD000612 [106], PXD009861 [107] и PXD001468 [108], каждый из которых был извлечён из общедоступной базы данных PRIDE. В ходе испытаний использовались данные по протеому человека, на основе которых были получены результаты поиска соответствующих спектров масс и последовательностей пептидов, причём отбор происходил на уровне ложноположительных обнаружений в пределах 1%. Эти результаты были

подвергнуты сравнению с показателями производительности таких инструментов, как Crux [109] и MSFragger [96, с. 1], с применением индивидуальных поисков в базах данных с использованием методов целевого отбора для определения точности FDR через инструмент Percolator [110]. Конкретные параметры тестирования включали строгий закрытый поиск, где допустимое отклонение массы прекурсора было ограничено до ±5 частей на миллион (ppm), что обеспечивает высокую точность согласования масс. Параметры фрагментации были зафиксированы на уровне 0,1 Да, так как спектры MS/MS предварительно преобразовывались в массивы с интервалами индексов, соответствующими диапазону масс в 0,1 Да. База данных белков для этих поисков была подготовлена с применением *in silico* триптического расщепления, при этом было разрешено наличие до двух пропущенных мест расщепления. Кроме того, размеры пептидов были ограничены диапазоном от 7 до 50 аминокислотных остатков, что позволяет учитывать большое разнообразие потенциальных пептидов, встречающихся в природе, и обеспечивает комплексный поиск по протеомным данным.

Для углубленного анализа протеомных данных были созданы специализированные базы данных пептидов, включающие версии без посттрансляционных модификаций, а также с одним и двумя потенциальными модификациями, такими как фосфорилирование и окисление на каждый пептид. Эти базы данных использовались для поиска в различных протеомных датасетах, собранных в результате экспериментальных исследований. Для создания таких баз данных были применены стратегии виртуального расщепления белков и добавления модификаций *in silico*, при этом применялись различные подходы к генерации фиктивных записей. Эти фиктивные записи получались путём перестановки аминокислотных последовательностей реальных пептидов, исключая их *N*- и *C*-концевые аминокислоты, чтобы предотвратить создание пептидов, которые могут существовать в природе и таким образом исключить возможность перекрытия между целевыми и фиктивными данными. Компиляция баз данных пептидов привела к созданию коллекций с огромным числом вариантов: приблизительно 2.6 миллиона пептидов без модификаций, 11.3 миллиона с одной модификацией фосфорилирования и 33.3 миллиона с двумя модификациями фосфорилирования. В то же время базы данных пептидов с окислением содержали около 2.5 миллиона пептидов для неокисленных вариантов, 3.7 миллиона для одиночных и 4.0 миллиона для двойных окисленных вариантов. Поиск пептидов проводился методом генерации теоретических масс-спектров с заранее установленными параметрами интенсивности. Для каждого спектра сообщались пять наилучших совпадений из каждой базы данных, включая как целевые, так и фиктивные варианты. С помощью таких инструментов, как Percolator и PeptideProphet, эти совпадения затем переупорядочивались для точной оценки частоты ложных срабатываний и среди них выбирались те, у которых оценка FDR оказывалась меньше 1%. Учитывая, что разработанная модель предназначена для генерации встраиваний пептидов, которые не зависят от их заряда, при сравнении производительности с такими программами, как Crux и MSFragger, было решено ограничить заряды

фрагментов теоретически генерируемых масс-спектров уровнем +1, чтобы обеспечить корректное сопоставление результатов. Визуальное сравнение эффективности различных методов поиска представлено в последующих иллюстрациях, где подробно демонстрируются результаты поиска по разным базам данных и оценки FDR, позволяя таким образом оценить и сравнить преимущества и недостатки каждого из подходов. По результатам сравнения можно сделать вывод о том, как различные стратегии обработки данных и алгоритмы поиска влияют на точность идентификации пептидов и общую производительность протеомного анализа. Важно отметить, что процесс валидации результатов с использованием разнообразных баз данных позволяет не только оценить способность алгоритмов корректно идентифицировать пептиды в условиях наличия или отсутствия посттрансляционных модификаций, но и демонстрирует важность использования строго контролируемых метрик для оценки FDR. Это подчёркивает необходимость точного и репрезентативного моделирования приманок, что является ключевым компонентом для обеспечения достоверности протеомных исследований. Результаты сравнения приведены на рисунках 8 и 9.

	SC_MS_Peptide_Ident		Crux		MSFragger	
	Пептиды	ПТМ	Пептиды	ПТМ	Пептиды	ПТМ
a	28.3	99.8	16.7	42.1	12.1	19.7
b	33.4	123.1	17.9	43.5	14.0	18.5

Рисунок 8 – Сравнение с инструментами с использованием набора данных NIST с одной модификацией фосфорилирования на пептид (*a*) и двумя модификациями фосфорилирования (*b*)

		SC_MS_Peptide_Ident		Crux		MSFragger	
		Пептиды	ПТМ	Пептиды	ПТМ	Пептиды	ПТМ
PXD000612	*a*	23.1	33.2	23.0	28.5	23.0	29.9
	b	23.2	34.8	23.2	28.7	22.0	27.6
	c	23.3	34.6	23.1	28.5	22.1	27.4
PXD009861	*a*	6.8	7.7	5.4	8.1	5.0	8.0
	b	7.8	8.3	7.3	9.2	6.0	9.1
	c	7.7	8.2	7.5	8.2	6.4	9.1
PXD001468	*a*	6.1	7.9	5.2	7.0	5.3	7.2
	b	5.9	7.9	5.2	7.1	5.3	7.3
	c	6.3	9.7	5.3	7.7	5.9	8.8

Рисунок 9 – Сравнение с инструментами с использованием наборов данных из базы данных PRIDE на не модифицированных пептидах (*a*), с одной модификацией фосфорилирования на пептид (*b*) и двумя модификациями фосфорилирования (*c*)

Исследование PXD000612 проводилось на основе тщательно курированной базы данных пептидов, отражающей широкий спектр посттрансляционных модификаций протеома человека, включая нефосфорилированные пептиды и те, которые подвергались фосфорилированию на серине, треонине и тирозине, для каждого из трёх подразделений: левого, центрального и правого сегментов соответственно. В контрасте с этим, анализ набора данных PXD009861 был направлен на исследование пептидов с различными уровнями окисления метионина, вплоть до двух модифицированных остатков на пептид. Отдельно, поиск для набора данных PXD001468 осуществлялся в базе данных, специально ориентированной на идентификацию пептидов с одним концевым ацетилированием и до двух сайтов окисления, устанавливая лимит в максимум две модификации на пептид, что позволяет глубже погрузиться в анализ сложных модификаций. Исследуемая модель была использована для сравнения её эффективности с другими инструментами поиска, такими как Crux и MSFragger, при этом особое внимание уделялось их способности к идентификации PSM. Сопоставление результатов показало, что модель может не только конкурировать, но и превосходить данные

инструменты по количеству идентификаций PSM, демонстрируя при этом аналогичные или даже более высокие результаты в контексте идентифицированного количества пептидов. Такой подход подчёркивает важность комплексного анализа протеомных данных с учётом разнообразия посттрансляционных модификаций, включая, но не ограничиваясь фосфорилированием и окислением, для получения наиболее полной картины протеомных изменений.

Дополнительно, в рамках сравнительного анализа было проведено исследование пересечения идентификаций пептидов, выполненных с помощью исследуемой сети, Crux и MSFragger, что позволяет наглядно оценить уникальность и сходство между инструментами. При работе с набором данных PXD000612 все три инструмента сумели идентифицировать 19 291 пептид без модификаций, 18 350 пептидов с одной посттрансляционной модификацией и 18 437 пептидов с двумя посттрансляционными модификациями. Для набора данных PXD009861 инструменты идентифицировали 2 813 пептидов без модификаций, 3 117 пептидов с одной посттрансляционной модификацией и 3 020 пептидов с двумя посттрансляционными модификациями. Для набора данных PXD001468 было обнаружено 3 018 пептидов без модификаций, 2 938 пептидов с одной посттрансляционной модификацией и 3 223 пептида с двумя посттрансляционными модификациями.

Исследование показывает, что большая часть пептидов, обнаруженных в ходе каждого эксперимента, были распознаны всеми тремя инструментами поиска, подчёркивая их способность к нахождению общих целевых последовательностей. Однако исследуемая сеть выявила значительно большее количество уникальных пептидов по сравнению с Crux и MSFragger, что свидетельствует о потенциальной эффективности этого инструмента в идентификации меньших по размеру пептидов, в среднем около десяти аминокислот в длину. Это может указывать на особенности алгоритма, которые оптимизированы для поиска и распознавания коротких пептидных последовательностей. Тем не менее, пептиды, обнаруженные исключительно с помощью Crux и MSFragger и упущенные оптимизированной моделью, демонстрируют аналогичное распределение по длине среди общей массы идентифицированных пептидов. Предположительно, SC_MS_Peptide_Ident быть более чувствительна к коротким пептидам, но Crux и MSFragger обладают общей способностью идентифицировать пептиды разнообразной длины.

Что касается заряда предшественника уникальных пептидов, здесь наблюдается согласованность между инструментами: большинство идентифицированных пептидов в каждом случае имели более низкий заряд. Это согласуется с ожиданиями, поскольку пептиды с более низким зарядом предшественника часто легче идентифицировать из-за меньшего количества возможных заряженных состояний, которые нужно рассмотреть во время масс-спектрометрического анализа. Напротив, Crux и MSFragger, при использовании настроек по умолчанию, которые включают в себя возможность распознавания фрагментов с зарядами выше +1, демонстрируют способность к выявлению большего количества уникальных пептидов с более высоким зарядом

предшественника. Это различие в настройках по умолчанию может влиять на итоговую статистику идентификации, давая приоритет пептидам с определёнными физико-химическими характеристиками.

Следует отметить, что предпочтение более низкозаряженных пептидов может быть связано с биологическими особенностями протеома, где более низкозаряженные пептиды обычно более распространены, или с ограничениями масс-спектрометрического оборудования и методик анализа данных. В то же время, распознавание пептидов с более высоким зарядом предшественника может быть критически важным для идентификации более длинных и сложных пептидов, которые могут играть ключевые роли в биологических процессах.

С учётом описанных наблюдений становится очевидным, что выбор инструмента для поиска пептидов может значительно влиять на результаты исследования протеома. Отсюда следует, что для более полного понимания комплексности протеома и эффективного выявления широкого спектра пептидов различной длины и заряда необходим комплексный подход, включающий использование нескольких инструментов и сравнение их результатов. Это также подчёркивает необходимость тщательной настройки инструментов в соответствии с целями исследования и характеристиками анализируемых протеомных данных.

Расширенная и оптимизированная сеть SpeCollate представляет собой мощный инструмент для протеомного анализа, который использует техники машинного обучения для проецирования спектров и пептидов в одно и то же векторное пространство. Это встраивание позволяет прямое сравнение спектров и пептидов без традиционных метрик подсчёта совпадений и симуляции масс-спектров, облегчая процесс идентификации и анализа. Используя передовые алгоритмы глубокого обучения, сеть демонстрирует потенциал для превосходства над классическими подходами в области идентификации пептидов. Методика, основанная на принципах кросс-модального обучения, внедряет новаторские подходы в анализ масс-спектрометрических данных. В результате обширных экспериментальных исследований, включая тестирование с использованием как заранее составленных спектральных библиотек, так и реальных масс-спектрометрических данных, было показано, что подходы кросс-модального обучения способны обеспечивать высокую точность идентификации пептидов. Эти техники глубокого обучения могут не только улучшать результаты по сравнению с традиционными методами, но и обнаруживать новые закономерности в данных, которые ранее оставались незамеченными. Эффективность кросс-модального обучения особенно заметна при анализе сложных спектров, где стандартные подходы могут столкнуться с трудностями из-за высокой степени шума или перекрывающихся пиков. Ключевым преимуществом исследуемой модели является способность сети изучать и адаптироваться к разнообразным особенностям спектральных данных, что делает её особенно подходящей для работы со сложными биологическими образцами и многомерными данными. Благодаря тому, что сеть обучается на реальных и обширных наборах данных MC/MC, она способна формировать обобщенные векторные представления, которые отражают не только массу и

интенсивность пиков, но и более тонкие свойства, такие как возможные посттрансляционные модификации и различные заряды ионов.

Интеграция кросс-модального обучения с существующими спектральными библиотеками усиливает эффективность процесса идентификации, позволяя сравнивать образцы с обширной базой данных известных спектров и пептидов. Это позволяет не только подтверждать наличие уже известных молекул в образце, но и открывать новые молекулы, расширяя тем самым понимание протеома.

Выводы по разделу

1. Процесс сопоставления масс-спектрометрических данных с соответствующими пептидными последовательностями играет центральную роль в сложной задаче определения пептидных структур в рамках протеомных исследований. Масс-спектрометрические спектры представляются комплексами измерений, каждое из которых фиксирует массу и соответствующую ей интенсивность ионов, причем эти данные обычно выражены в формате с плавающей точкой, что позволяет зафиксировать высокую степень точности масс ионов. В свою очередь, пептиды описываются через последовательности аминокислот, где каждая аминокислота обозначается определённой буквой в алфавите аминокислот, формируя тем самым уникальные буквенные цепочки, соответствующие биологическим молекулам. Эта буквенная нотация пептидов служит биохимическим языком, который должен быть "переведён" в числовые данные спектров для верификации и анализа. Задача сопоставления, таким образом, связывает качественные и количественные аспекты биологических исследований, требуя использования сложных вычислительных методов для преобразования и интерпретации каждого сигнала из спектра в конкретную аминокислотную последовательность. Это сопоставление не только выявляет, какие пептиды присутствуют в образце, но и предоставляет информацию о возможных модификациях, таких как фосфорилирование или окисление, которые могут оказывать влияние на функцию и динамику белков в клетке.

2. На основании общедоступной сети подобия SpeCollate был разработан метод SC_MS_Peptide_Ident для изучения и оценки сходства между пептидами и их масс-спектрами, объединяя в себе две различные формы данных: числовые значения масс-спектрометрических спектров и последовательности аминокислот, кодируемые алфавитным обозначением. Специально разработанная спектральная подсеть, которая включает два полностью связанных слоя, предназначена для тщательного анализа и обработки многомерных спектральных данных. В рамках этой сети спектральная подсеть с двумя полностью связанными слоями отвечает за обработку спектральных данных, тогда как пептидная подсеть, построенная на базе двунаправленной LSTM архитектуры, дополненной ещё двумя полностью связанными слоями, используется для анализа последовательностей аминокислот. Взаимодействие этих подсетей в рамках единой структуры позволяет проводить сложные сравнения и распознавать тонкие закономерности сходства между пептидами и

спектрами. Ключевую роль в оптимизации процесса обучения играет функция потерь SNAP-loss, которая ориентирована на логистические потери.

3. Алгоритм SC_MS_Peptide_Ident прошёл процесс обучения в течение 50 эпох, в результате которого была достигнута точность проверки на уровне 94%. Этот показатель свидетельствует о высоком уровне адаптации модели к задачам идентификации пептидов по масс-спектрам. После завершения этапа обучения модель подверглась тестированию на четырёх различных наборах данных, что позволило оценить её производительность в сравнении с широко используемыми инструментами идентификации в данной области – Crux и MSFragger. В ходе сравнительного анализа для обеспечения справедливости сравнения был установлен специфический параметр для Crux и MSFragger: заряд фрагментов теоретических спектров, генерируемых этими инструментами, ограничивался значением в +1. Это условие позволило установить одинаковые начальные условия для всех инструментов, чтобы исключить влияние данного параметра на результаты идентификации. При таком условии предложенная SC_MS_Peptide_Ident продемонстрировал более высокие показатели точности верных идентификаций.

3 РАЗРАБОТКА АЛГОРИТМА ПРЕДСКАЗАНИЯ ФУНКЦИЙ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ПРИМЕНЕНИЕМ ВОЗМОЖНОСТЕЙ МАШИННОГО ОБУЧЕНИЯ

3.1 Описание, анализ и предобработка исходных наборов данных

В качестве основных источников данных в экспериментальной части исследования использовались общедоступные базы данных Pfam, UniProtKB/Swiss-Prot и Gene Ontology. В диссертационном исследовании были использованы наборы данных, полученные напрямую из указанных источников, а также тот же набор данных, что и в [32, с. 161], и наборы данных, использованные в [111] и [112]. Эксперименты по обучения моделей проводились независимо на разных наборах данных.

Pfam представляет собой обширную базу данных, хранящую информацию о семействах белковых доменов. В этой базе каждое семейство характеризуется через множественное выравнивание белковых последовательностей, а также сопутствующую скрытую марковскую модель, которая служит для идентификации новых членов семейства в различных последовательностях. Отмечено, что значительная доля последовательностей в базе данных UniProtKB, конкретно 77.2% из приблизительно 137 миллионов, ассоциируется по крайней мере с одной аннотацией из Pfam, что указывает на широкое покрытие и полезность этой ресурсной базы в белковой аннотации. По состоянию на ноябрь 2021 года, база данных Pfam содержала 19 632 уникальные записи, организованные в 657 кланов, каждый из которых объединяет семейства с общими эволюционными предками.

В базе данных Pfam каждая запись имеет несколько ключевых атрибутов, которые вместе предоставляют целостную информацию о белковых доменах:

- sequence: это поле включает в себя аминокислотную последовательность, которая специфична для идентифицированного домена в составе белка, и не относится к полной последовательности белка.
- family_accession: уникальный идентификатор или номер доступа, который служит для определения и отслеживания конкретного белкового семейства в Pfam.
- sequence_name: название или обозначение последовательности, которое может быть использовано для поиска или справки.
- align_sequence: последовательность, представленная в контексте множественного выравнивания, что позволяет увидеть её взаимосвязь с другими схожими последовательностями.
- family_id: обозначение семейства, выраженное обычно в виде лаконичного названия, отражающего его функциональные или эволюционные особенности.

В таблице 3 показаны некоторые примеры используемых данных из базы данных Pfam.

Таблица 3 – Примеры данных из базы данных Pfam

№	Наименование поля	Данные
1	sequence:	VLERKISTRQTREELIKKGVLIPD
	family_accession	PF02755.15
	sequence_name	I3IWL9_ORENI/33-56
	aligned_sequence	VLERKISTRQTREELIKKGVLIPD
	family_id	RPEL
2	sequence	NPCTIDSCGPKGCVHIAMSCDDN
	family_accession	PF00526.18
	sequence_name	F0ZFD3_DICPU/581-603
	aligned_sequence	NPCTIDSC.GPK....G.CVHIAM.SCDDN
	family_id	Dicty_CTDC
3	sequence:	KLNSLGGLVALNLGSIDNASASGTLV
	family_accession	PF07581.12
	sequence_name:	Q7WYX3_PSEAI/240-265
	aligned_sequence	KLNSLGGLVALNL.....GSIDNASASG.TLV
	family_id	Glug

В первой части исследования по обучению модели для предсказания функций белков использовались только данные из Pfam. В типичном эталонном тесте для оценки алгоритмов машинного обучения база данных Pfam, содержащая 17 929 семейств белковых доменов, подверглась процедуре случайного разбиения на три набора данных: один для обучения модели, другой для настройки (валидации) параметров модели и последний для проверки производительности модели. Согласно этому разбиению, большинство последовательностей (80%) используется в обучающем наборе, обеспечивая модели достаточное количество данных для извлечения и обобщения характеристик, в то время как оставшиеся 20% делятся поровну для настройки и тестирования модели, что позволяет провести оценку точности и обобщающей способности модели на данных, которые не использовались во время обучения. На рисунке 10 показано общее количество последовательностей в наборе данных.

```
[ ]  # Given data size
     print('Train size: ', len(df_train))
     print('Val size: ', len(df_val))
     print('Test size: ', len(df_test))
```

```
Train size:  1086741
Val size:  126171
Test size:  126171
```

Рисунок 10 – Общее количество последовательностей в каждом наборе данных из базы данных Pfam

Таким образом, для построения и обучения модели использовались три набора данных: набор обучающих данных *train*, набор данных *val* и набор тестовых данных *test*, состоящие из 1 086 741 последовательностей, 126 171 последовательностей и 126 171 последовательностей соответственно.

Для общей статистической обработки данных было посчитано количество кодов (аминокислот) в каждой невыровненной последовательности, выявлено, что большинство несогласованных аминокислотных последовательностей имеют количество символов в диапазоне 50-250, результат изображён на рисунке 11.

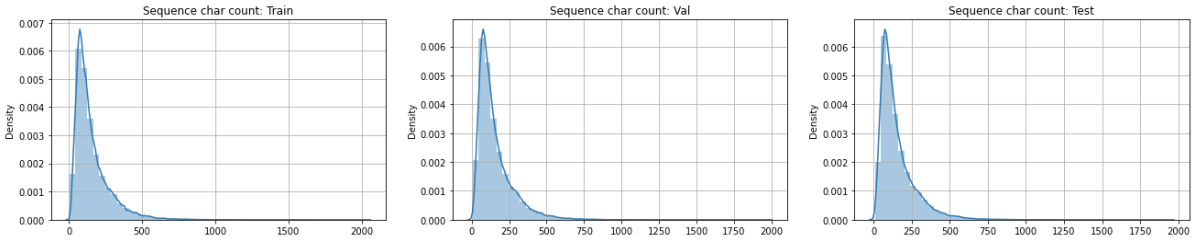


Рисунок 11 – Количество символов последовательности

После группировки данных по family_id, было рассчитано, какие семейства чаще всего встречаются в выборке, как показано на рисунке 12.

```
family_id
Methyltransf_25    3637
LRR_1              1927
Acetyltransf_7     1761
His_kinase         1537
Bac_transf         1528
Lum_binding        1504
DNA_binding_1      1345
Chromate_transp    1265
Lipase_GDSL_2      1252
DnaJ_CXXCXGXG      1210
SRP54_N            1185
WD40               1173
OTCace_N           1171
PEP-utilizers      1147
Glycos_trans_3N    1138
THF_DHG_CYH        1113
Prenyltransf       1104
HTH_1              1064
Maf                1061
DHH                1057
```

Рисунок 12 – Наиболее часто встречающиеся семейства

Проведение подсчёта частоты появления для каждого кода (аминокислоты) в каждой невыровненной последовательности изображено на рисунке 13. Наиболее частым аминокислотным кодом является лейцин (L), за которым следуют аланин (A), валин (V) и глицин (G). Очевидно, что аминокислоты E, X, U, B, O и Z присутствуют в очень маленьком количестве.

Отсюда можно выбрать только 20 общих аминокислот для кодирования последовательности на этапе предварительной обработки.

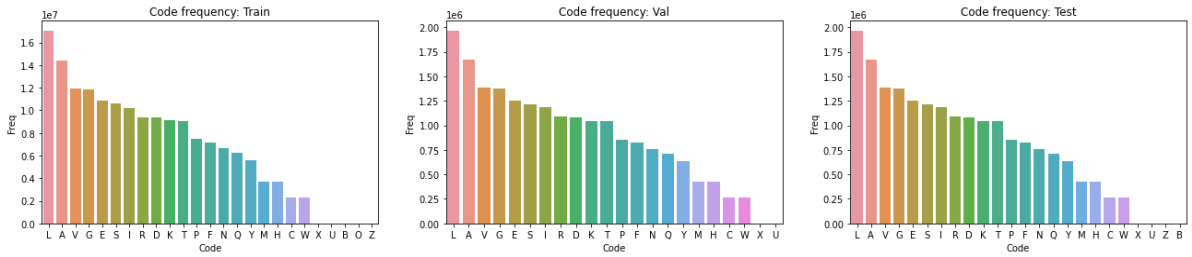


Рисунок 13 – Частота кода (аминокислоты) последовательности

Для трансформации белковых последовательностей в формат, подходящий для машинного обучения, используется методика присвоения уникальных целочисленных значений каждой аминокислоте, сопоставляя их с буквенными кодами в порядке возрастания их уникальных идентификаторов. Таким образом, аминокислоты кодируются цифрами, что позволяет компьютерным системам более эффективно обрабатывать биологические данные. Далее применяется "one-hot encoding" или метод горячего кодирования, где для каждого типа аминокислоты создаётся отдельный двоичный столбец (признак). В этом столбце устанавливается значение "1" для тех позиций в последовательности, где присутствует конкретная аминокислота, и "0" в противном случае. Этот подход позволяет эффективно представлять последовательности в виде векторов в многомерном пространстве, что исключительно полезно для компьютеризированных методов классификации и прогнозирования в различных биоинформатических приложениях, таких как предсказание функции белков. Аминокислотные последовательности представлены соответствующим 1-буквенным кодом, например, код для аланина (A), аргинина (R) и так далее.

Пример последовательности:

PHPESRIRLSTRRDAHGMPIPRIESRLGPDAFARLRFMARTCRAILAAAGCAAPFEEFSSA DAFSSTHVFGTCRMGHDPMRNVVDGWGRSHRWPNLFVADASLFPSSGGGESPGLTIQALALRT

Для каждой последовательности один буквенный код заменяется на число. После кодирования вышеупомянутая последовательность будет иметь такой вид:

[13, 7, 13, 4, 16, 15, 8, 15, 10, 16, 17, 15, 15, 3, 1, 7, 6, 11, 13, 8, 13, 15, 8, 4, 16, 15, 10, 6, 13, 3, 1, 5, 1, 15, 10, 15, 5, 11, 1, 15, 17, 2, 15, 1, 8, 10, 1, 1, 1, 6, 2, 1, 1, 13, 5, 4, 4, 5, 16, 16, 1, 3, 1, 5, 16, 16, 17, 7, 18, 5, 6, 17, 2, 15, 11, 6, 7, 3, 13, 11, 15, 12, 18, 18, 3, 6, 19, 6, 15, 16, 7, 15, 19, 13, 12, 10, 5, 18, 1, 3, 1, 16, 10, 5, 13, 16, 16, 6, 6, 6, 4, 16, 13, 6, 10, 17, 8, 14, 1, 10, 1, 10, 15, 17]

Для улучшения процесса обучения машинных алгоритмов в биоинформатике, иногда простых числовых векторов недостаточно, и требуется использование более сложных структур данных, например матриц. Этот процесс обычно включает в себя несколько ключевых этапов:

1. *Нормализация длины последовательности*: исходная аминокислотная последовательность приводится к стандартной длине, предположим, 1000 единиц, чтобы обеспечить одинаковый размер входных данных для обучения нейронной сети [113]. Это достигается за счёт добавления специальных символов (например, подчёркивания "_"), которые не встречаются среди стандартных аминокислот, для последовательностей, длина которых менее 1000 аминокислот.

2. *Создание таблицы соответствия между аминокислотами и числовыми кодами*: используя утвержденный список IUPAC [114] для аминокислот, каждой аминокислоте присваивается уникальный порядковый номер, который используется для их идентификации в последующем кодировании.

3. *Визуализация аминокислот в виде вектора в двумерном пространстве*: каждая аминокислота представляется точкой в этом пространстве, где ось *X* отражает порядковые номера аминокислот, а ось *Y* отражает их расположение в последовательности, как показано на рисунке 14. Неаминокислотные позиции в этой матрице обозначаются нулями, а аминокислотные позиции – единицами, создавая тем самым двумерное представление последовательности, которое может быть использовано для обучения и анализа в машинных алгоритмах классификации и предсказания.

Отсортированные аминокислотные коды (21)

Аминокислотная последовательность

	A	C	D	…	M	N	…	Y	W
M	0	0	0		1	0		0	0
B	0	0	0.5		0	0.5		0	0
A	1	0	0		0	0		0	0
C	0	1	0		0	0		0	0
…									
Y	0	0	0		0	0		1	0
W	0	0	0		0	0		0	1

Рисунок 14 – Представление последовательности в виде матрицы после кодирования исходной аминокислотной последовательности

Конечное представление обучающей последовательности выглядит следующим образом:

[4 9 13 10 20 16 5 7 12 7 8 1 1 18 9 1 8 17 19 16 13 7 14 15 6 8 10 1 16 6 6 6 17 1 3 15 17 8 9 10 19 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Для некоторых специфичных алгоритмов предобработка данных выглядит иным образом, но обобщённо все основные алгоритмы так или иначе берут за основу численное представление последовательности.

Для второй части экспериментальных исследований была выбрана база данных UniProtKB/Swiss-Prot [115] – информационный ресурс, являющийся частью комплексной базы данных UniProt и представляющий данные о белках, которые подвергаются строгому процессу верификации и проверки. Этот процесс выполняется квалифицированными специалистами в области биохимии и молекулярной биологии, которые используют как проверенные научные публикации, так и современные инструменты для обеспечения точности аминокислотных последовательностей и соответствующих функциональных аннотаций Gene Ontology для исследуемых белков. Эти аннотации GO считаются высококачественными и достоверными, поскольку базируются на данных, тщательно отобранных и проанализированных экспертами. Как представлялось ранее, Gene Ontology представляет собой широко применяемую систему стандартизации функциональной аннотации, которая включает в себя три основные подонтологии, отражающие разные аспекты белковой функции: биологические процессы, молекулярные функции и клеточный компонент. Эти категории позволяют систематизировать и детально описать разнообразные функции белков, их роли в жизнедеятельности клетки и организма. В результате каждый белок может быть аннотирован не одной, а множеством функциональных аннотаций GO, каждая из которых представляет собой отдельный аспект его биологической активности и участия в различных биологических процессах.

В данном эксперименте были использованы три набора данных, полученных из базы данных UniProtKB/Swiss-Prot: *Maze*, *Indica* и *Japonica*. Процесс извлечения и фильтрации данных, а также их оригинальное представление в источнике продемонстрированы на рисунке 15.

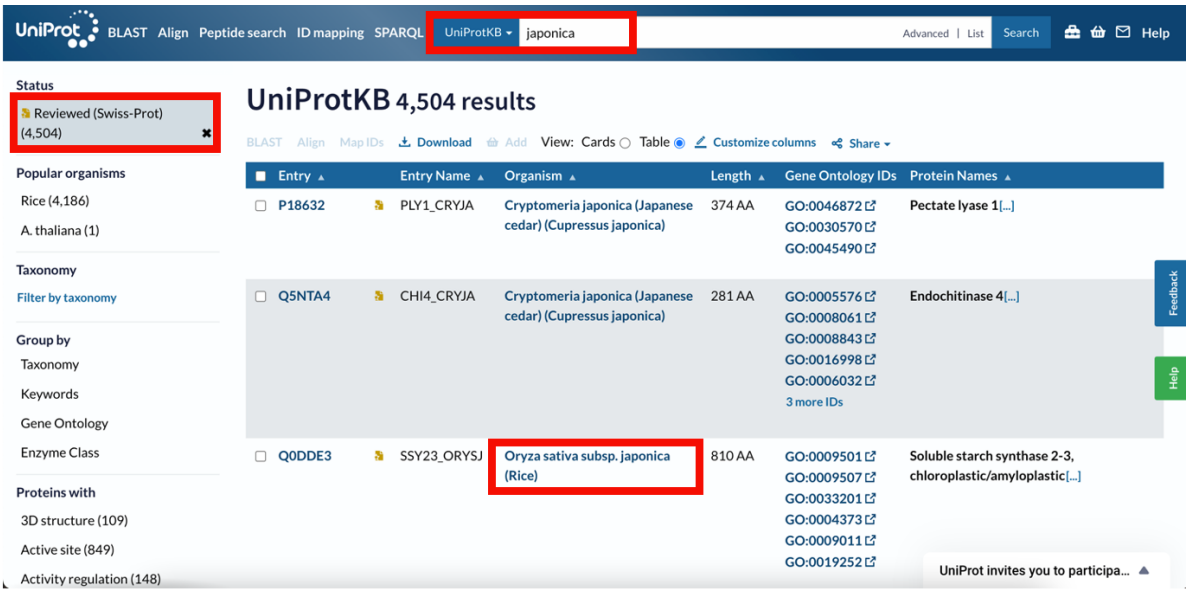


Рисунок 15 – Получение данных из UniProtKB/Swiss-Prot

Полученные файлы содержат различную информацию о белках выбранного организма. Информация об аминокислотной последовательности и присвоенных экспертами терминах GO представлена в формате, показанном на рисунке 16.

```
SQ   SEQUENCE   460 AA;  48500 MW;  3EDE49A203C8E4CC CRC64;
     MWDLNDSPAA EAAPPPLSPS ADDSGASSSS AAAVVEIPDD ADDDSAAVVV VTRQFFPPAV
     PGGGGDPAPG NARAGWLRLA GAAPPVAATG PAASAAVSKK SRRGPRSRSS QYRGVTFYRR
     TGRWESHIWD CGKQVYLGGF DTAHAAARAY DRAAIKFRGV EADINFSLED YEDDLKQMSN
     LTKEEFVHVL RRQSTGFPRG SSKYRGVTLH KCGRWEARMG QFLGKKYVYL GLFDTEEEAA
     RAYDRAAIKC NGKDAVTNFD PSIYAGEFEP PAAATGDAAE HNLDLSLGSS AGSKRGNVDG
     GGDDEITGGG GGGAGSDQRV PMAFDLDWQT AAARSTKAKF DQNSNHPQMP PVLQVTHLPF
     SPRHHHQFLS NGDPGTAGGL SLTIGAGMAG HWPPQQQQGW GNAGGMSWPH PPHPPPPPTN
     AAAAATATAA AASSRFPPYI ATQASTWLQK NGFHSLTRPT
//
ID   LRSK7_ORYSJ             Reviewed;         695 AA.
AC   A0A0P0VIP0; Q0E1D9; Q6K3K2;
DT   12-AUG-2020, integrated into UniProtKB/Swiss-Prot.
DT   20-JAN-2016, sequence version 1.
DT   24-JAN-2024, entry version 47.
DE   RecName: Full=L-type lectin-domain containing receptor kinase S.7 {ECO:0000305};
DE            Short=OsLecRK-S.7 {ECO:0000303|PubMed:31833176};
DE            EC=2.7.11.1 {ECO:0000269|PubMed:31833176};
DE   AltName: Full=Protein DEFECTIVE IN APERTURE FORMATION 1 {ECO:0000303|PubMed:32284546};
DR   ExpressionAtlas; A0A0N7KJT8; baseline and differential.
DR   GO; GO:0005634; C:nucleus; IDA:UniProtKB.
DR   GO; GO:0005667; C:transcription regulator complex; IDA:UniProtKB.
DR   GO; GO:0003700; F:DNA-binding transcription factor activity; IDA:UniProtKB.
DR   GO; GO:0000976; F:transcription cis-regulatory region binding; IDA:UniProtKB.
DR   GO; GO:0060860; P:regulation of floral organ abscission; IMP:UniProtKB.
DR   GO; GO:0009909; P:regulation of flower development; IMP:UniProtKB.
DR   GO; GO:0080050; P:regulation of seed development; IMP:UniProtKB.
DR   GO; GO:0009409; P:response to cold; IEP:UniProtKB.
DR   GO; GO:0097548; P:seed abscission; IMP:UniProtKB.
DR   CDD; cd00018; AP2; 1.
```

Рисунок 16 – Образец представления данных о белках в формате UniProtKB/Swiss-Prot

Для определения содержимого белковых аннотаций в эксперименте отбирались только те GO-термины, которые были подкреплены экспериментальными данными, обозначенными кодами EXP, IDA, PI, IMP, IGI, TEP, TAS и IC:

```
selected_experiments = ['EXP', 'IDA', 'IPI', 'IMP', 'IGI', 'IEP', 'TAS', 'IC']
df_filtered = df[df['Evidence'].isin(selected_experiments)].drop_duplicates()
```

Белки, которые не имели никаких GO-аннотаций, исключались из анализа. С целью уточнения категории аннотации функции GO, данные были организованы в отдельные наборы данных в соответствии с каждой из трёх подонтологий MF, BP и CC. В процессе распределения аннотаций использовалась иерархическая структура GO, обеспечивающая, чтобы если белок был аннотирован конкретной GO-аннотацией, он также наследовал все соответствующие аннотации, связанные с его "родительскими" терминами в иерархии. Это гарантировало, что белки с разными подонтологиями, например,

те, которые имели аннотации как в BP, так и в MF, классифицировались соответственно в обоих наборах данных.

После всех этих шагов было вычислено количество белков, аннотированных для каждого класса GO, для целей прогнозирования модели были отобраны все классы, имеющие 50 или более аннотаций (в разных экспериментах количество аннотаций варьировалась). Далее, сформированные данные были случайно разделены на обучающую (80%) и тестовую (20%) выборки, а обучающая выборка в дальнейшем была разделена на обучающую (80%) и валидационную (20%) подвыборки. На финальном этапе обработки все наборы данных были преобразованы методом горячего кодирования, описанным ранее.

3.2 Реализация и обучение модели машинного обучения BiLSTM

Разработанный метод в первую очередь включает использование двунаправленной сети долговременной краткосрочной памяти (BiLSTM), что позволяет извлекать информацию как из общего контекста, так и из конкретных локальных свойств белков [116]. Этот подход расширяет границы возможностей извлечения данных о последовательности, превосходя ограничения, связанные с традиционными однонаправленными моделями, и сохраняя при этом целостность информационной последовательности, обеспечивая улучшенную точность в прогнозировании.

Кроме того, для углубления понимания взаимосвязи между различными атрибутами последовательности и для акцентирования на наиболее значимых аспектах последовательности, применяется механизм самовнимания (self-attention). Этот механизм позволяет модели адаптивно фокусироваться на ключевых особенностях последовательности, выделяя важные функциональные участки и повышая тем самым общую производительность, точность и применимость предсказательной модели в разнообразных условиях и на различных наборах данных. Подобные инновации в методиках машинного обучения значительно усиливают потенциал автоматизированного предсказания функций белков, способствуя более глубокому и системному пониманию их роли в биологических системах.

BiLSTM – это один из типов рекуррентных нейронных сетей, который обрабатывает данные последовательности как в прямом, так и в обратном направлении с двумя отдельными скрытыми слоями. BiLSTM основан на вентилях ввода, забывания и вывода. Для расчёта прогнозных значений используются следующие формулы (20) [117]:

$$\begin{gathered}\text{input gate}(i_t) = \sigma_g(W_i X_t + R_i h_{t-1} + b_i),\\ \text{forget gate}(f_t) = \sigma_g\left(W_f X_t + R_f h_{t-1} + b_f\right),\\ \text{cell candidate}(c_t) = \sigma_g(W_c X_t + R_c h_{t-1} + b_c),\end{gathered} \tag{20}$$

$$\text{output gate}(o_t) = \sigma_g(W_o X_t + R_o h_{t-1} + b_o),$$

где σ_g – функция активации вентиля, а W_i, W_f, W_c, и W_o входные весовые матрицы, тогда как R_i, R_f, R_c, и R_o – весовые матрицы, соединяющие предыдущее выходное состояние ячейки с тремя вентилями и входное состояние ячейки. X_t – вход, и h_{t-1} выход в предыдущий момент времени (t–1). b_i, b_f, b_o и b_c – векторы смещения.

На каждой временной итерации t состояние выхода ячейки C_t, и выход слоя h_t можно рассчитать следующим образом [118]:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{21}$$

$$h_t = o_t * \tanh(C_t) \tag{22}$$

Архитектура двунаправленной модели LSTM представлена на рисунке 17.

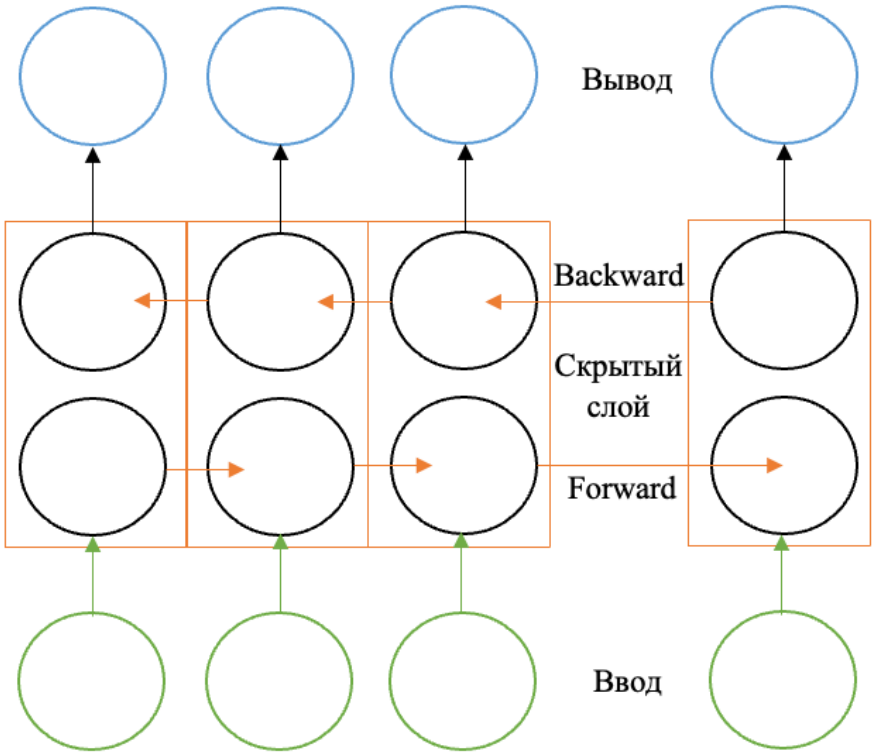


Рисунок 17 – BiLSTM-архитектура

Использование двунаправленного LSTM запускает ввод двумя способами, позволяя сохранять контекстную информацию из прошлого и будущего в любой момент времени. Алгоритм BiLSTM может собирать важную информацию об аминокислотных последовательностях в двух направлениях, полностью учитывать информацию о контекстуальной корреляции текущих

аминокислотных последовательностей и может более глубоко изучать особенности белковых последовательностей.

Однако из-за длинной аминокислотной последовательности модель BiLSTM не может уловить самую прямую связь между вектором признаков и меткой результата. Добавление в модель механизма self-attention [119] может решить эту проблему. Он может взвешивать входные функции и измерять важность каждой функции для экспериментального объекта. Механизм self-attention широко используется в области классификации текстов и изображений [120, 121] машинного перевода [122] и биоинформатики [123]. В этой экспериментальной модели взаимосвязь вычислений в механизме self-attention показана на рисунке 18.

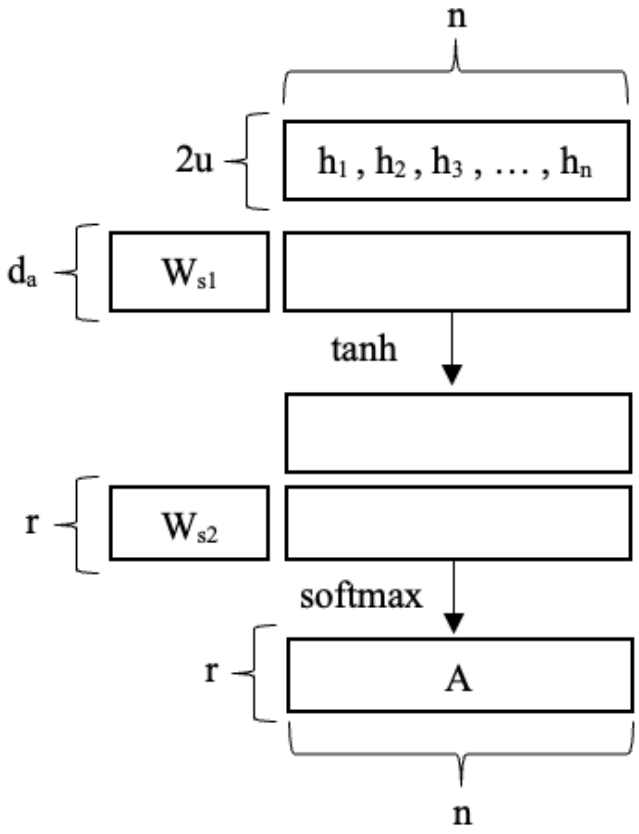


Рисунок 18 – Взаимосвязь вычислений в механизме self-attention

Для признаков, находящихся далеко друг от друга и взаимозависимых, требуется определённое количество времени и шагов, чтобы накопить достаточно информации, чтобы связать их. Чем дальше они друг от друга, тем меньше вероятность того, что сеть BiLSTM захватит эффективную информацию. Это означает, что, когда аминокислота может быть связана с окружающими ее аминокислотами или более отдаленными аминокислотами, использование алгоритма BiLSTM учитывает только информацию до и после последовательности белка в определённом диапазоне и не может решить проблему корреляции между прерывистыми аминокислотами. Стоит отметить,

что одна аминокислота или несколько аминокислот могут иметь большое влияние на функцию белка. В процессе расчёта механизм self-attention может напрямую связать корреляцию между любыми двумя функциями в последовательности за один шаг расчёта, что значительно сокращает расстояние между зависимыми функциями на большом расстоянии. Интегрирование механизма self-attention в структуру BiLSTM привносит новый уровень аналитических способностей, позволяя модели адаптивно выделять и фокусироваться на наиболее значимых аминокислотах в последовательности. Этот аспект является фундаментальным для выявления потенциально активных сайтов и критически важных областей в белковой молекуле, которые могут иметь первостепенное значение для ее функционирования или взаимодействия с другими молекулами.

Модель последовательно вводит каждый элемент входной последовательности X_t в сеть BiLSTM. Затем рассчитываются прямой выход h_t и обратный выход h_t в прямом и обратном направлениях соответственно. После этого выходные векторы как в прямом, так и в обратном направлении суммируются и может быть получен окончательный выходной вектор h_t. Далее вектор признаков $H = (h_1, h_2, \ldots, h_n)$, полученный с помощью модели BiLSTM, вводится в слой самовнимания для расчёта весового вектора a. Выражение весового вектора a можно получить следующим образом:

$$a = softmax\ (W_{s2} \tanh(W_{s1} H^T)), \tag{23}$$

где W_{s2} и W_{s1} являются матрицами параметров, размерность весового вектора a равна n, размерность вектора m равна $2u$. Перемножив вектор признаков и вектор весов a, можно получить окончательный вектор m слоя самовнимания.

Затем вектор m, выводимый из слоя самовнимания, передаётся в полносвязный слой. Наконец, выходные данные отображаются в диапазоне [0, 1] через сигмовидную функцию слоя активации, тем самым получая результат прогнозирования функции белка. Архитектура алгоритма PFP_SelfAttn_BiLSTM показана на рисунке 19.

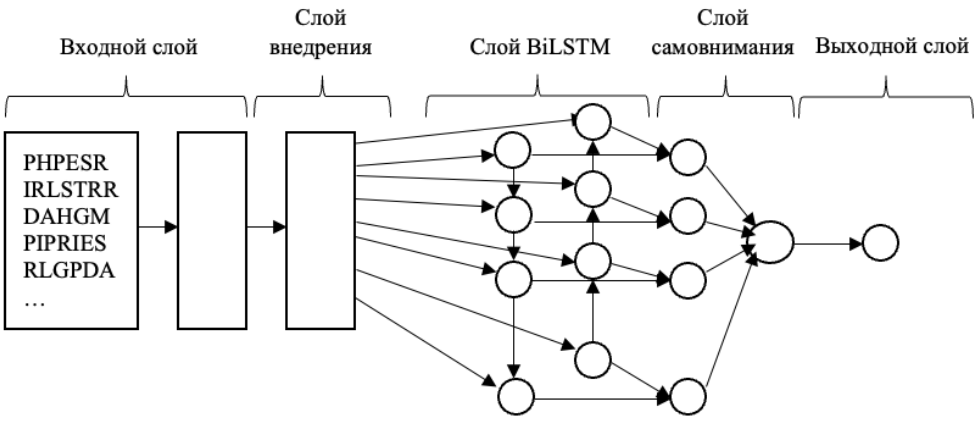


Рисунок 19 – Архитектура Self-attention – BiLSTM

Архитектура модели состоит из слоя внедрения для изучения векторного представления для каждого кода, за которым следуют BiLSTM и слой self-attention. Чтобы предотвратить переобучение, в модель добавлен слой Dropout. Выходной слой даёт значения вероятности для всех уникальных классов, и на основе наибольшей прогнозируемой вероятности модель классифицирует последовательности аминокислот к одному из белков семейства.

Техническая реализация были исполнена на языке Python с использованием библиотек TensorFlow, Keras, Keras-self-attention, Biopython и Goatools и запущена на облачной платформе Colab (https://colab.research.google.com/). Нейронная сеть была реализована следующим образом:

```
model = Sequential([
Masking(mask_value=0., input_shape=(max_length - gram_len + 1, len(vocab) + 7)),
Bidirectional(LSTM(64, return_sequences=True)),
Bidirectional(LSTM(32, return_sequences=True)),
SeqSelfAttention(attention_activation='sigmoid'),
FlattenAttentionLayer(attention_dim=32),
Dropout(0.35),
Dense(256, activation='relu'),
Dropout(0.35),
Dense(num_classes, activation='sigmoid')
])
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

В процессе разработки модели машинного обучения были задействованы следующие ключевые компоненты:

− для оптимизации параметров модели применялся алгоритм Adam, который является эффективным методом стохастической оптимизации и предоставляет автоматические корректировки скорости обучения для каждого параметра;

− в качестве критерия оценки производительности модели использовалась метрика Accuracy, отражающая процент правильно классифицированных образцов и служащая в качестве целевой функции для оптимизации;

− функция потерь бинарной кросс-энтропии применялась для вычисления ошибок в предсказаниях модели. Эта функция оценивает вероятностные прогнозы относительно истинных значений, предоставляя меру расхождения между распределениями вероятности, предсказанными моделью, и фактическими данными, что позволяет модели учиться от своих ошибок и повышать точность классификации:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log(\hat{y}_i) + (1 - \hat{y}_i) * \log(1 - \hat{y}_i) \quad (24)$$

В контексте многоклассовой классификации каждый объект имеет свою истинную метку класса, обозначаемую как y_i, которая соответствует фактическому классу, к которому он принадлежит. С другой стороны, предсказание модели, то есть метка класса, предсказанная классификатором для этого же объекта, обозначается как $\hat{y}_i$. Задача модели – максимально точно приблизить $\hat{y}_i$ к y_i для каждого объекта в наборе данных. Количество различных классов в наборе данных обозначается как *N*. Это число важно, поскольку оно определяет размерность выходного вектора в задачах классификации, где каждый класс представлен уникальным выходным узлом в модели.

Бинарная кросс-энтропия, используемая в контексте логистической регрессии, применяется как способ количественной оценки степени ошибок в бинарных классификациях. Она вычисляет значение потерь (*Loss*), основываясь на разнице между ожидаемыми и реальными исходами классификации. Эта функция потерь особенно полезна, когда модель должна выводить вероятности принадлежности к классам, так как позволяет преобразовать выходные данные модели в вероятности и использовать их для расчёта степени неуверенности модели в её предсказаниях.

3.3 Результаты разработки и обучения модели для предсказания функций белков

Для обучения модели в разных экспериментах были запущены от 10 до 50 эпох обучения с размерностью серий от 32 до 64. В начальных экспериментах модель обучалась на данных базы Pfam. Для предотвращения переобучения модели были выставлены следующие параметры:

earlystopping = EarlyStopping(monitor='val_loss', patience=3, verbose=1).

В Таблице 4 представлены результаты обучения PFP_SelfAttn_BiLSTM сети на наборе данных Pfam.

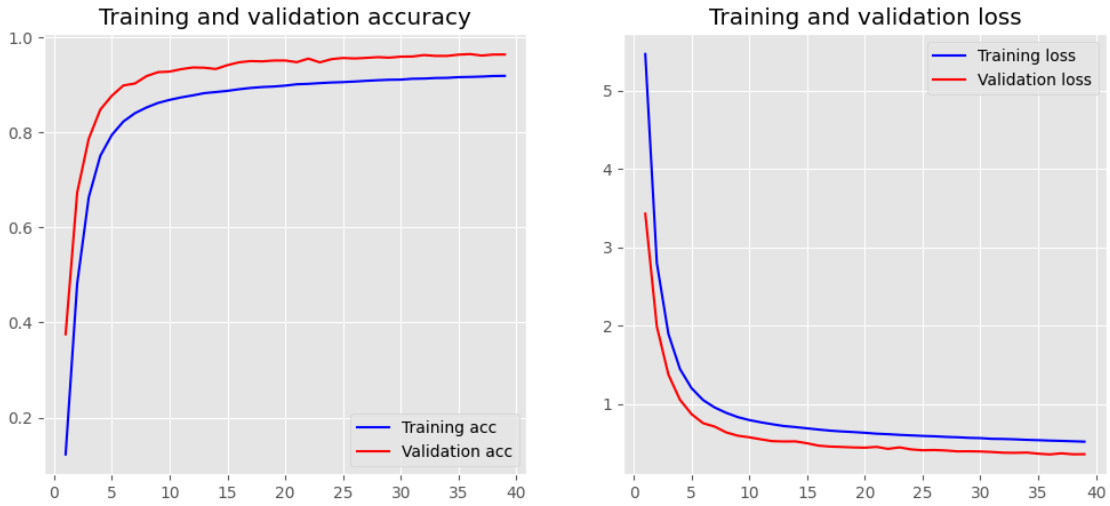
Таблица 4 – Результаты обучения нейронной сети PFP_SelfAttn_BiLSTM на наборе данных Pfam

Epoch	Loss	Accuracy	Val_loss	Val_accuracy
1	2	3	4	5
1	5.46	0.12	3.43	0.37
2	2.80	0.48	1.99	0.67
3	1.89	0.66	1.37	0.78
4	1.44	0.75	1.05	0.84
5	1.20	0.79	0.87	0.87
…				
46	0.53	0.91	0.36	0.96
47	0.53	0.91	0.35	0.96
48	0.53	0.91	0.37	0.96

Продолжение таблицы 4

1	2	3	4	5
49	0.52	0.91	0.35	0.96
50	0.52	0.91	0.36	0.96

Рисунок 20 представляет собой детальный график обучения, на котором демонстрируется динамика обучения двунаправленной нейронной сети с долгосрочной кратковременной памятью в течение 50 эпох обучения. На первом графике можно наблюдать увеличение показателя точности (Accuracy) модели, который отражает процент правильно классифицированных примеров в обучающем наборе данных на каждом этапе обучения. Этот показатель является ключевым для оценки способности модели правильно распознавать и ассоциировать входные данные с соответствующими категориями. Параллельно, второй график отображает уменьшение функции потерь (Loss), которая является математическим выражением расхождения между предсказаниями модели и реальными данными. Обычно это значение снижается по мере того, как модель находит более оптимальные параметры в процессе обучения, тем самым уменьшая разницу между предсказанным и фактическим результатами. Именно снижение значения Loss указывает на то, что модель становится всё более точной и надёжной в своих предсказаниях.



a b

Рисунок 20 – Результаты обучения модели PFP_SelfAttn_BiLSTM на наборе данных Pfam по показателю accuracy: *a* – значение accuracy, *b*– значение функции потерь

На рисунке 21 отражен конечный результат значения функции потерь и показателя Accuracy для набора данных Pfam.

```
display_model_score(model1,
    [train_pad, y_train],
    [val_pad, y_val],
    [test_pad, y_test],
    256)
```

```
1717/1717 [==============================] - 9s 5ms/step - loss: 0.3380 - accuracy: 0.9699
Train loss:  0.338043212890625
Train accuracy:  0.9699085354804993
----------------------------------------------------------------------
213/213 [==============================] - 1s 5ms/step - loss: 0.3611 - accuracy: 0.9639
Val loss:  0.3611033260822296
Val accuracy:  0.9638640880584717
----------------------------------------------------------------------
213/213 [==============================] - 1s 5ms/step - loss: 0.3630 - accuracy: 0.9632
Test loss:  0.363040953874588
Test accuracy:  0.9632388353347778
```

Рисунок 21 – Результат обучения модели PFP_SelfAttn_BiLSTM на наборе данных Pfam

Оценка производительности разработанной модели на данных базы Pfam показала, что наилучшая точность BiLSTM составила 0.9699 для обучающего набора, 0.9638 для набора данных и 0.9632 для тестового набора.

В следующем эксперименте модель обучалась на трёх наборах данных, полученных из базы данных UniProtKB/Swiss-Prot: *Maze*, *Indica* и *Japonica.* Для обучения модели на каждом из тестовых наборов данных были выставлены следующие параметры:

```
callbacks_list = [
    ModelCheckpoint('model.h5', monitor='val_loss', save_best_only=True, verbose=1),
    CSVLogger('training_log.csv', append=True)
]

history = model.fit(
    X_train, Y_train,
    epochs=20,
    batch_size=32,
    validation_data=(X_valid, Y_valid),
    callbacks=callbacks_list,
    verbose=2
).
```

Обучение модели проводилось по десять раз для каждой подонтологии Gene Ontology, на рисунке 22 изображены точность и потери при обучении в одном из запусков обучения нейронной сети для организма *Indica* подонтологии клеточный компонент.

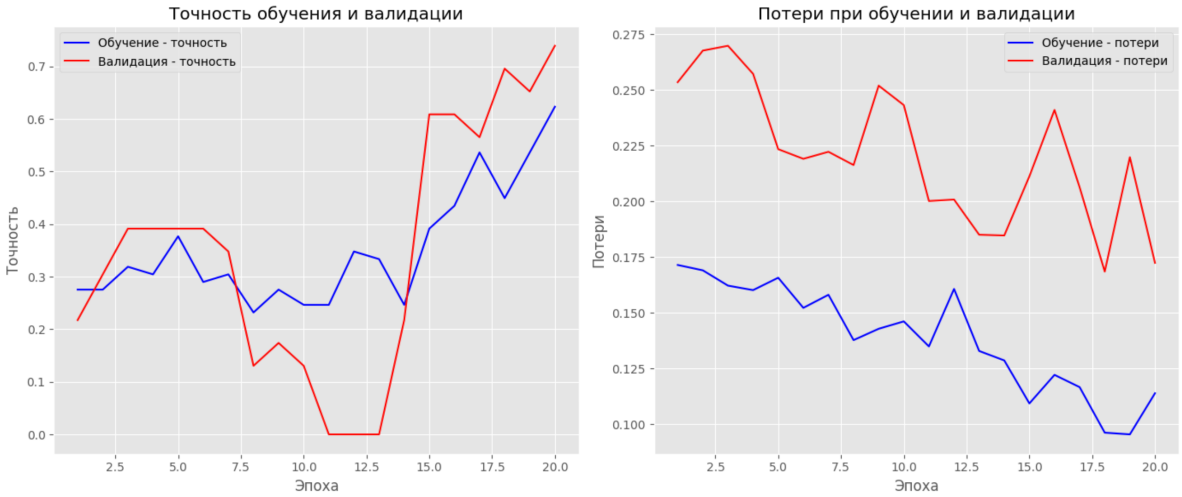


Рисунок 22 – Результаты обучения модели PFP_SelfAttn_BiLSTM на наборе данных *Indica* по показателю accuracy: *a* – значение accuracy, *b*– значение функции потерь

В таблице 5 представлены результаты обучения на том же наборе данных для нескольких запусков.

Таблица 5 – Результаты обучения PFP_SelfAttn_BiLSTM на наборе данных *Indica*

Набор данных	Подтип онтологии	№ запуска	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Indica	КК	1	0.66	0.88	0.89	0.93
		2	0.75	0.91	0.99	0.95
		3	0.89	0.91	0.99	0.95
		…				
		9	0.87	0.91	0.97	0.94
		10	0.89	0.92	0.99	0.95

В результате обучения модели удалось достигнуть хороших результатов по всем основным показателям, продемонстрировав точность, достигающую показателя 0.91. Несмотря на очевидность того, что показатели стали несколько ниже по сравнению с результатами экспериментов на наборе данных Pfam, это объясняется намного меньшим объёмом обучающей выборки в экспериментах с данными UniProtKB/Swiss-Prot. Отсюда выводится важное наблюдение, что при увеличении количества данных PFP_SelfAttn_BiLSTM показывает лучшие результаты, но теряет стабильность показателя Accuracy при большом числе запусков обучения модели со средним количеством эпох обучения 20.

3.4 Оценка точности идентификации исследуемого метода

При оценке производительности алгоритмов прогнозирования функций белков принято использовать четыре стандартизированных метрики, которые обеспечивают комплексный анализ качества предсказаний модели [61, с. 46]. Эти метрики включают:

- Чувствительность (SE), также известная как recall или истинно положительная ставка, отражает способность алгоритма корректно идентифицировать положительные случаи. Это соотношение истинно положительных результатов (*TP*) к общему числу фактических положительных случаев, которое включает *TP* и ложноотрицательные (*FN*) результаты.

- Специфичность (SP), иногда называемая истинно отрицательной ставкой, показывает эффективность алгоритма в правильном исключении отрицательных случаев и рассчитывается как отношение истинно отрицательных результатов (*TN*) к сумме *TN* и ложноположительных (*FP*) результатов.

- Точность (ACC) измеряет общую долю правильных предсказаний, независимо от класса, и вычисляется как доля суммы *TP* и *TN* к общему числу всех предсказаний, включая все четыре категории *TP*, *TN*, *FP* и *FN*.

- Коэффициент корреляции Мэтьюса (MCC) представляет собой балансированную меру качества алгоритма, учитывающую все четыре переменные *TP*, *TN*, *FP* и *FN* в предсказаниях. MCC является одной из самых надёжных метрик в условиях несбалансированных классов и может принимать значения от -1 до 1, где 1 указывает на идеальное предсказание, 0 – на не лучше случайного предсказания, и -1 – на полное несоответствие между предсказанием и реальностью.

Важно отметить, что каждая из этих метрик даёт уникальную информацию о производительности и их совместное рассмотрение обеспечивает полное понимание сильных и слабых сторон предиктора. Эффективность алгоритма прогнозирования не должна оцениваться по одной только метрике, так как разные аспекты производительности могут быть лучше выявлены через разные метрики. Таким образом, комплексный анализ всех четырёх показателей является ключевым для достижения наиболее объективной и всесторонней оценки эффективности алгоритмов прогнозирования функции белка.

В частности, SE определяется процентом истинно положительных образцов, правильно идентифицированных как «положительные» (25):

$$SE = \frac{TP}{TP+FN} \tag{25}$$

SP указывает долю истинно отрицательных образцов, которые были правильно предсказаны как «отрицательные» (26):

$$SP = \frac{TN}{TN+FP} \tag{26}$$

ACC относится к количеству истинных образцов (положительных плюс отрицательных), делённому на количество всех изученных образцов (27):

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{27}$$

MCC (F1) можно рассчитать по следующей формуле (28):

$$MCC = \frac{(TP*TN-FP*FN)}{\sqrt{(TP+FN)*(TP+FP)*(TN+FP)*(TN+FN)}} \quad (28)$$

В таблице 6 показаны значения всех четырёх показателей для разработанной модели на примере нескольких тестовых наборов данных, также приведены результаты запуска нескольких классических моделей на тех же наборах данных.

Экспериментальные результаты алгоритма PFP_SelfAttn_BiLSTM сравниваются с алгоритмами CNN, LSTM и BLSTM. Эксперименты были проведены по десять раз для каждой подонтологии и выведено среднее значение результатов всех десяти экспериментов.

Таблица 6 – Результаты предсказания функций белков на тестовых данных

Набор данных	Подтип онтологии	Алгоритм	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	2	3	4	5	6	7
Maize	БП	CNN	0.77	0.73	0.62	0.67
		LSTM	0.79	0.75	0.67	0.70
		BiLSTM	0.77	0.69	0.67	0.68
		PFP_SelfAttn_BiLSTM	**0.80**	**0.74**	**0.69**	**0.72**
	МФ	CNN	0.82	0.64	0.73	0.69
		LSTM	0.83	0.70	0.71	0.70
		BiLSTM	0.83	0.68	0.69	0.69
		PFP_SelfAttn_BiLSTM	**0.86**	**0.74**	**0.75**	**0.74**
	КК	CNN	0.78	0.79	0.61	0.68
		LSTM	0.82	0.79	0.70	0.74
		BiLSTM	0.78	0.78	0.63	0.69
		PFP_SelfAttn_BiLSTM	**0.83**	**0.79**	**0.75**	**0.77**
Indica	БП	CNN	0.73	0.50	0.88	0.64
		LSTM	0.89	0.81	0.77	0.79
		BiLSTM	0.87	0.78	0.74	0.76
		PFP_SelfAttn_BiLSTM	**0.90**	**0.83**	**0.85**	**0.81**
	МФ	CNN	0.84	0.73	0.51	0.60
		LSTM	0.86	0.66	0.61	0.63
		BiLSTM	0.84	0.63	0.63	0.63
		PFP_SelfAttn_BiLSTM	**0.87**	**0.75**	**0.71**	**0.68**
	КК	CNN	0.81	0.78	0.77	0.77
		LSTM	0.82	0.80	0.77	0.78
		BiLSTM	0.81	0.78	0.77	0.78
		PFP_SelfAttn_BiLSTM	**0.85**	**0.91**	**0.89**	**0.95**
Japonica	БП	CNN	0.78	0.71	0.58	0.64
		LSTM	0.80	0.71	0.63	0.67
		BiLSTM	0.79	0.65	0.66	0.65
		PFP_SelfAttn_BiLSTM	**0.81**	**0.76**	**0.64**	**0.68**
	МФ	CNN	0.83	0.75	0.85	0.80
		LSTM	0.87	0.82	0.88	0.85
		BiLSTM	0.84	0.78	0.85	0.81

Продолжение таблицы 6

1	2	3	4	5	6	7
		PFP_SelfAttn_BiLSTM	**0.87**	**0.84**	**0.88**	**0.86**
	КК	CNN	0.78	0.75	0.78	0.76
		LSTM	0.81	0.81	0.80	0.80
		BiLSTM	0.79	0.74	0.82	0.77
		PFP_SelfAttn_BiLSTM	**0.83**	**0.87**	**0.84**	**0.89**

Из экспериментальных результатов видно, что по сравнению с алгоритмами CNN, LSTM и BiLSTM алгоритм PFP_SelfAttn_BiLSTM имеет самую высокие показатели accuracy, F1 и precision. Это доказывает возможность применения алгоритма PFP_SelfAttn_BiLSTM для прогнозирования функций других белков, а также закладывает основу для применения алгоритма PFP_SelfAttn_BiLSTM для прогнозирования общих функций белков в будущем.

Далее внимание уделяется сегменту данных, где обнаружены расхождения в прогнозируемой функции белка. Сопоставление предсказанной функции белка проводится на основе верификации с актуальной информацией, зафиксированной в базе данных Swiss-Prot. Анализ неверных предсказаний особенно важен для выявления потенциальных улучшений в прогнозных моделях. Ошибочные предсказания могут указывать на специфические проблемы в обучающих данных, методах предсказания или в самой структуре алгоритмов. Результаты сравнения функций белков, отобранных в рамках данного анализа, подробно представлены в таблице 7, которая включает в себя статистику по каждому случаю расхождения, а также предполагаемые причины ошибок и возможные пути их коррекции. Эта информация может служить отправной точкой для последующих итераций усовершенствования методов машинного обучения, применяемых в PFP_SelfAttn_BiLSTM для автоматизированной функциональной аннотации белков.

Таблица 7 – примеры предсказания функций белков *Indica* и *Japonica*

Набор данных	Подтип онтологии	Белок	Функция	Предсказанная функция
Indica	КК	Q01N44	GO:0005783	GO:0005783
			GO:0005886	GO:0005886
			GO:0016020	GO:0016020
			GO:0005789	GO:0005789
				GO:0009536
Japonica	БП	Q9DE67	GO:0030199	GO:0030199
			GO:0007601	GO:0007601
			GO:0032914	GO:0032914
			GO:0045944	

На последнем этапе исследования был проведён анализ результативности прогнозирования белковых функций на примере *Indica*. По данным базы Swiss-Prot, белок Q01N44, известный также как FAAH_ORYSI, ассоциирован с двумя

функциональными ролями: GO:0016042 и GO:0006629. Функция GO:0016042 детально описывает механизмы расщепления липидов, обозначая специфические химические реакции и метаболические пути, ответственные за катаболизм этих биомолекул. С другой стороны, GO:0006629 устанавливает более общий процесс липидного метаболизма, включающий все химические превращения, затрагивающие липиды. В рамках данного эксперимента прогнозирование функций указанного белка не ограничивается только этими двумя функциями, но также предполагает наличие связанной функции GO:0006807, характеризующейся как метаболический процесс, включающий соединения, содержащие азот – в этом случае прогноз неожиданно расширил спектр потенциальных биологических активностей белка, указывая на возможное участие в азотосодержащих соединениях. Такое расширение прогнозируемых функций, вероятно, объясняется перекрёстным метаболическим взаимодействием между липидным обменом и метаболизмом азота, что подчёркивает сложность и многоаспектность белковых функций и необходимость более детального анализа таких взаимосвязей для повышения точности предсказательных моделей.

Анализ функциональных ролей белка E0ZS48, также известного как UREA_ORYSI, выявил его связь с четырьмя ключевыми функциональными доменами, описанными в GO: GO:0009039, GO:0016787, GO:0016810 и GO:0046872. В частности, функция GO:0009039 указывает на специфическую роль уреазы, фермента, отвечающего за расщепление уреи. Гидролазная активность, обозначаемая GO:0016787, включает в себя широкий спектр ферментативных функций, способных катализировать гидролиз многочисленных видов молекулярных связей, в то время как GO:0016810 конкретизирует класс ферментов, действующих на разнообразные углерод-азотные связи за исключением пептидных. Далее, GO:0046872 подчёркивает важность функции белка в связывании с ионами металлов, что имеет значение во множестве биохимических путей и структурных конфигураций. Тем не менее, экспериментальное тестирование указывает на сложности в прогнозировании GO:0016810. Эта проблема может быть обусловлена тесным функциональным перекрытием между GO:0016910 и GO:0016787, поскольку обе эти категории ферментов участвуют в процессах гидролиза, но с различной специфичностью относительно типов связей, что создаёт трудности для алгоритмов предсказания в разграничении между этими двумя видами активности.

Комплексный анализ функциональных атрибутов белка FAAH_ORYSI (Q01N44) выявил его причастность к различным структурным компонентам клетки, включая эндоплазматический ретикулум (GO:0005783), плазматическую мембрану (GO:0005886), клеточную мембрану (GO:0016020), и мембрану эндоплазматического ретикулума (GO:0005789). Эти функции отражают ключевую роль белка в поддержании структурной целостности и функциональной активности этих критически важных клеточных структур. В рамках проведённого предсказательного анализа, помимо точного предсказания этих четырёх функций, была идентифицирована дополнительная функциональная роль, связанная с пластидами (GO:0009536), являющимися

частью семейства органелл, встречающихся в клетках растений и некоторых простейших, и содержащих мембрано-связанную ДНК. Как правило, пластиды включают в себя широкий спектр органелл, таких как митохондрии и центросомы, играющие важную роль в энергетическом метаболизме и клеточной регуляции. Предсказание функции пластиды GO:0009536, по всей видимости, обусловлено её связью с мембранной системой клетки, в том числе с эндоплазматическим ретикулумом. Такая связь может указывать на участие белка в процессах, связанных с функционированием и взаимодействием различных мембранных и мембрано-связанных структур в клетке. Данное предсказание подчёркивает необходимость дополнительного изучения взаимодействий белков с разнообразными клеточными компонентами и потенциальную взаимозависимость функций различных мембранных органелл.

Белок Q9DE67 (LUM_COTJA), как подтверждено данными из базы данных Swiss-Prot, участвует в нескольких ключевых биологических процессах, определяемых функциональными терминами Генной Онтологии: коллагеновая склеротизация (GO:0030199), восприятие света (GO:0007601), положительная регуляция транскрипции фактором РНК-полимеразы II (GO:0045944) и положительное регулирование продукции трансформирующего фактора роста бета (GO:0032914). Эти функции указывают на сложную роль белка в регуляции разнообразных физиологических и развивающихся процессов в организме. Тем не менее, в ходе предсказательного анализа было обнаружено, что функция GO:0045944, обозначающая стимуляцию активности транскрипции фактором РНК-полимеразы II, не была предсказана. Это отсутствие может быть связано с тонкими нюансами регуляторных путей, участвующих в этом процессе, поскольку функция GO:0032914 также указывает на стимуляцию производства важного цитокина, трансформирующего фактора роста бета. Обе функции имеют общую черту – положительное регулирование биохимических путей, что может приводить к перекрывающимся или взаимно связанным путям в клетке и, как следствие, к сложности в их точном прогнозировании через компьютерные алгоритмы.

При исследовании молекулярных функций растительного белка Q7XFK2 (BGA14_ORYS) организма *Japonica*, было выявлено четыре основные аннотации, соответствующие карбогидразной активности (GO:0030246), гидролазной активности (GO:0016787), гликозилгидролазной активности (GO:0016798) и каталитической активности (GO:0003824). Сравнение с данными из Swiss-Prot показало, что все эти функции, за исключением каталитической активности, были аннотированы ранее. В данном эксперименте функция каталитической активности (GO:0003824) была определена как дополнительная, предполагая участие белка в каталитическом процессе биохимических реакций при физиологических температурах. Учитывая, что гидролазная активность (GO:0016787) и гликозилгидролазная активность (GO:0016798) включают в себя катализ разрыва определённых химических связей, они тесно связаны с более общей категорией каталитической активности (GO:0003824).

Для белка Q7XWK5 (SAG39_ORYS) указаны четыре основные функциональные аннотации. Термин GO:0005615 относится к

экстрацеллюлярному пространству, области вне клеточной мембраны, где происходит множество важных взаимодействий между клетками и молекулами. GO:0005764 описывает лизосому, специализированную клеточную органеллу, ответственную за разложение макромолекул. Дополнительно, функции GO:0010282 и GO:0005773 ассоциируются с вакуолярными структурами, выполняющими различные хранилищные и транспортные функции в клетке.

Интересно отметить, что в процессе прогнозирования для белка Q7XWK5 была обнаружена дополнительная функция GO:0005634, относящаяся к ядру клетки. Эта органелла представляет собой мембрано-ограниченную структуру, являющуюся местом размещения и репликации генетического материала организма. Присутствие этой функции в результатах прогнозирования может отражать биологическую родственность между вакуолярными органеллами и ядром, учитывая, что обе структуры являются отсеками, заключенными в мембраны, и выполняют жизненно важные функции для поддержания клеточной жизнедеятельности и наследственности.

Таким образом, алгоритм PFP_SelfAttn_BiLSTM, хотя и не гарантирует безошибочное предсказание всех функций для белков с близкими или перекрывающимися функциями, в целом продемонстрировал значительную точность в прогнозировании функций белков. Обзор результатов в таблицах 6 и 7 подчёркивает успешность алгоритма в достижении достоверных прогнозов. Важно отметить, что помимо верификации уже известных функций белков, представленных в базе данных Swiss-Prot, алгоритм также способен расширять существующие знания, предсказывая новые функциональные аннотации GO, отсутствующие в текущих записях, что даёт новое направление для последующих экспериментальных исследований.

Выводы по разделу

1. Основными источниками данных для обучения представленной модели были выбраны база данных Pfam, содержащая обширную коллекцию информации о семействах белковых доменов, и база данных UniProtKB-Swiss-Prot, хранящая информацию о белках и курируемая экспертами. В рамках подготовительного этапа исследования была проведена тщательная предварительная обработка исходных данных с целью исключения возможных ошибок и несоответствий, что позволило повысить качество и эффективность последующего обучения модели. Для кодирования белковых последовательностей был выбран метод горячего кодирования.

2. Ядром модели служит комбинация BiLSTM и механизма self-attention, которая устраняет проблемы, связанные с корреляцией разрозненных аминокислотных остатков, возникающих в процессе биосинтеза белков. Архитектура модели состоит из слоя внедрения для изучения векторного представления для каждого кода, за которым следуют BiLSTM и слой self-attention, дополненные слоем Dropout для предотвращения переобучения. Эта многоуровневая структура поддерживает глубокое исследование скрытых паттернов в последовательностях и обеспечивает высокую точность в предсказании функций белков.

3. Модель обучалась на разных наборах данных в течение 10-50 эпох с размерностью серий от 32 до 64 и достигла точности в 96%. Для тестирования модели были выбраны два набора данных из открытых источников. Результаты предсказания функций для белков организмов *Indica* и *Japonica* были оценены при помощи ручного аннотирования с помощью международной базы данных Swiss-Prot. Тесты показали хорошие результаты предсказания функций белков с разными функциями, однако дали менее точные результаты при аннотировании очень схожих функций белков. Это подчёркивает необходимость улучшения алгоритмических решений и включения дополнительных параметров, которые могут повысить способность предсказательной модели различать близкие функциональные категории, особенно когда они вовлечены в похожие биохимические процессы.

ЗАКЛЮЧЕНИЕ

В диссертационном исследовании был выполнен анализ современных решений для работы как с данными масс-спектрометрии, так и с данными белковых последовательностей, находящихся в открытых и общедоступных международных базах данных. На основании проведённого анализа текущего состояния описанных биологических баз данных было отмечено несоответствие потребностей научного сообщества и возможностей доступного инструментария для проведения всесторонних исследований экспоненциально растущих данных в таких базах как GenBank, UniProt, PFam и другие. Были обозначены две открытые проблемы:

1. Методы поиска пептидов в базах данных позволяют выявлять пептиды, превращая теоретически возможные пептиды в соответствующие им теоретические масс-спектры и сравнивая их с реально полученными экспериментальными спектрами. Для осуществления этого сравнения применяются различные эвристические функции схожести, которые предназначены для определения наилучшего совпадения между теорией и экспериментом. Однако этот подход имеет определённые ограничения: эвристический подход к подсчёту баллов зачастую не учитывает все факторы, влияющие на схожесть, а простота преобразования пептидных последовательностей в теоретические спектры не всегда обеспечивает достаточную точность, особенно при анализе зашумлённых экспериментальных данных. Всё это обуславливает важность создания новой модели машинного обучения, способной на более глубоком уровне изучить и понять функциональные связи между экспериментальными спектрами и пептидными последовательностями, на основе аннотированных данных масс-спектрометрических исследований, обладающей способностью к преодолению указанных недостатков традиционного подхода и обеспечивающей повышенную точность идентификации. В свою очередь, методы секвенирования *de novo* ориентированы на прямое определение аминокислотной последовательности пептидов из экспериментальных масс-спектров без прямого обращения к базам данных. Это достигается благодаря анализу массовых различий между фрагментными ионами, что позволяет восстановить первичную структуру пептида. Основное преимущество этого подхода заключается в его способности распознавать новые или модифицированные пептиды, которые отсутствуют в базах данных. Алгоритмы *de novo* часто используют сложные модели и вычислительные стратегии для поиска оптимальных путей в графах фрагментаций, что требует значительных вычислительных ресурсов и времени. Таким образом, несмотря на их потенциал, существует необходимость в улучшении этих методов для повышения их точности и эффективности, чтобы они могли лучше справляться с широким спектром спектрометрических данных и обеспечивать более точное определение неизвестных пептидов и белков, включая те, которые не могут быть обнаружены традиционными подходами поиска по базе данных.

2. Идентификация функциональных особенностей белков путём ручной аннотации, несмотря на признание в научном сообществе как золотого стандарта, обеспечивающего высокую точность данных, сталкивается с ограничениями в виде значительных финансовых и временных затрат, что делает его сложным в масштабировании для больших геномных проектов. В свете растущего объёма геномных данных появилась потребность в автоматизации процесса аннотации, что привело к разработке автоматизированных методов. Эти методы нацелены на эффективную обработку обширных наборов данных, одновременно улучшая качество аннотации путём исключения человеческого фактора. Однако, существующие методы прогнозирования функций белков, основанные на традиционных моделях, часто не в состоянии уловить сложные нелинейные зависимости между структурными характеристиками белков и их биологическими функциями, отражёнными в терминах Gene Ontology. Это подчёркивает необходимость внедрения методов глубокого обучения, способных выявлять скрытые зависимости в биологических данных, которые традиционные алгоритмы машинного обучения не могли обнаружить из-за огромного объёма, избыточности и шума в белковых последовательностях, тем самым обеспечивая более точные и надежные прогнозы функций белка на основе доступной геномной информации.

Для решения вышеописанных проблем в ходе диссертационного исследования были решены поставленные задачи и получены следующие результаты :

1. Путём расширения и оптимизации общедоступной сети глубокого подобия с открытым кодом SpeCollate реализован алгоритм SC_MS_Peptide_Ident для изучения кросс-модальной функции сходства между пептидами и спектрами для выявления совпадений пептид-спектр. Модель обеспечивает обучение функции сходства, которая встраивает данные спектров и пептидов в одно евклидово пространство для их непосредственного сравнения. Чтобы улучшить качество совпадений, SC_MS_Peptide_Ident использует функцию потерь SNAP-loss, обучающуюся на массиве данных, состоящем из секстиплетов точек, для уменьшения расстояний между совпадающими парами и увеличения расстояний между несовпадающими. Модель показала высокую точность тестирования в 94%. Проведено сравнение эффективности алгоритма SC_MS_Peptide_Ident с существующими инструментами для идентификации пептидов.

2. Реализован алгоритм для классификации белковых последовательностей, позволяющий идентифицировать функции белков и присваивать им термины Gene Ontology. В основе алгоритма лежит двунаправленная сеть LSTM, скомбинированная с механизмом самовнимания, что позволяет решить ситуацию, когда аминокислота может быть связана с окружающими её аминокислотами или более отдаленными аминокислотами и использование только алгоритма BiLSTM не может решить проблему корреляции между прерывистыми аминокислотами. Механизм self-attention напрямую связывает корреляцию между любыми двумя функциями в последовательности за один шаг расчёта, что значительно сокращает расстояние

между зависимыми функциями на большом расстоянии. Отсюда можно утверждать, что комбинация алгоритма BiLSTM и self-attention уделяет больше внимания аминокислотам, которые могут иметь большое влияние на функцию белка, так как аминокислотная последовательность вносит больший вклад в точное предсказание функции белка. На основании проведённой ручной аннотации тестовых данных было экспериментально доказано, что применение алгоритма PFP_SelfAttn_BiLSTM показало достаточно высокую точность функционального аннотирования белковых последовательностей.

Таким образом, можно сделать вывод о том, то цель диссертационного исследования достигнута:

- предложен алгоритм для идентификации пептидных последовательностей, полученных на основании экспериментальных данных масс-спектрометрии;
- предложен алгоритм для определения функций неаннотированных белковых последовательностей на основе машинного обучения;
- модели реализованы на языке Python, обучены на данных из открытых источников, оценены с помощью общепринятых методов оценки (Accuracy, Loss), корректность предсказанных данных проверена путём сравнения с результатами ручного аннотирования.

Методики и алгоритмы, предложенные в рамках диссертационного исследования, могут быть применены к различным наборам биологических данных и демонстрировать хорошие результаты при правильной и качественной обработке входящих наборов данных.

Следующим этапом развития темы диссертации является, во-первых, расширение объёма входящих данных и оптимизация процесса предварительной обработки данных для обучения моделей, а также расширение моделей машинного обучения, что приведёт к ещё более качественным предсказаниям, во-вторых, создание графического интерфейса, позволяющего использовать разработанные модули в удобном для конечного пользователя формате, и, в-третьих, расширение функционала алгоритмов для вывода дополнительной информации об исследуемых белках и пептидов путём использования различных биологических баз данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Gstaiger M., Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics, and network biology // Nature Reviews Genetics. – 2009 – Vol. 10. – P. 617-627.

2 Perkins D.N., Pappin D.J., Creasy D.M., Cottrell J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data // Electrophoresis. – 1999 – Vol. 20. – P. 3551-3567.

3 Eng J.K., McCormack A.L., Yates J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database // Journal of the American Society for Mass Spectrometry. – 1994. – Vol. 5. – P. 976-989.

4 Craig R., Beavis R.C. TANDEM: matching proteins with tandem mass spectra // Bioinformatics. – 2004. – Vol. 20. – P. 1466-1467.

5 Geer L.Y., Markey S.P., Kowalak J.A., Wagner L., Xu M., Maynard D.M., et al. Open mass spectrometry search algorithm // Journal of Proteome Research. – 2004. – Vol. 3. – P. 958-964.

6 Li D., Fu Y., Sun R., Ling C.X., Wei Y., Zhou H., et al. pFind: a novel database- searching software system for automated peptide and protein identification via tandem mass spectrometry // Bioinformatics. – 2005. – Vol. 21. – P. 3049-3050.

7 Villalba G.C., Matte U. Fantastic databases and where to find them: Web applications for researchers in a rush // Genetics and Molecular Biology. – 2021. – Vol. 44.

8 Consortium GO. Gene ontology consortium: going forward // Nucleic Acids Research. – 2015. – Vol. 43. – P. 1049-1056

9 Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry // Rapid Communications in Mass Spectrometry. – 2003. – Vol. 17. – P. 2337-2342.

10 Chi H. et. al. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra // Journal of Proteome Research. – 2013. – Vol. 12. – P. 615-625.

11 Tran N.H., Zhang X., Xin L., Shan B., Li M. De novo peptide sequencing by deep learning // Proceedings of the National Academy of Sciences. – 2017. – Vol. 1141. – P. 8247-8252.

12 Brosch M., Yu L., Hubbard T., Choudhary J. Accurate and sensitive peptide identification with Mascot Percolator // Journal of Proteome Research. – 2009. – Vol. 8. – P. 176-181.

13 Geer L.Y., Markey S.P., Kowalak J.A., Wagner L., Xu M., Maynard D.M., Yang X., Shi W., Bryant S.H. Open mass spectrometry search algorithm // Journal of Proteome Research. – 2004. – Vol. 5. – P. 958-964.

14 Zhang J., Xin L., Shan B., Chen W., Xie M., Yuen D., Zhang W., Zhang Z., Lajoie G.A., Ma B. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification // Molecular & Cellular Proteomics. – 2012. – Vol. 11

15 Kim S., Mischerikow N., Bandeira N., Navarro J.D., Wich L., Mohammed S., Heck A.J., Pevzner P.A. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search // Molecular & Cellular Proteomics. – 2010. – Vol. 12. – P. 2840-2852.

16 MacLean B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments // Bioinformatics. – 2010. – Vol. 26. – P. 966-968.

17 Lam H., Aebersold R. Spectral library searching for peptide identification via tandem MS // Methods in Molecular Biology. – 2010. – Vol. 604. – P. 95-103.

18 Piovesan D., Giollo M., Leonardi E., Ferrari C., Tosatto S.C. Inga: protein function prediction combining interaction networks, domain assignments and sequence similarity // Nucleic Acids Research. – 2015. – Vol. 43. – P. 134-140.

19 Gong Q., Ning W., Tian W. Gofdr: a sequence alignment-based method for predicting protein functions // Methods. – 2022. – Vol. 93. – P. 3-14.

20 Cozzetto D., Minneci F., Currant H., Jones D.T. Ffpred 3: feature-based function prediction for all gene ontology domains // Scientific Reports. – 2016. – Vol. 6. – P. 1-11.

21 Jung J., Yi G., Sukno S.A., Thon M.R. Pogo: prediction of gene ontology terms for fungal proteins // BMC Bioinformatics. – 2010. – Vol. 11. – P. 215.

22 You R., Huang X., Zhu S. Deeptext2go: improving large-scale protein function prediction with deep semantic text representation // Methods. – 2018a. – Vol. 145. – P. 82-90.

23 You R., Yao S., Xiong Y., Huang X., Sun F., Mamitsuka H., Zhu S. Netgo: improving large-scale protein function prediction with massive network information // Nucleic Acids Research. – 2019. – Vol. 47. – P. 379-387.

24 Rifaioglu A.S., Dogan T., Martin M.J., Cetin-Atalay R., Atalay V. Deepred: automated protein function prediction with multi-task feed-forward deep neural networks // Scientific Reports. – 2019. – Vol. 9. – P. 1-16

25 Zhang F., Song H., Zeng M., Wu F-X., Li Y., Pan Y., Li M. A deep learning framework for gene ontology annotation with sequence-and network-based information // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2020.

26 Kulmanov M., Hoehndorf R. Deepgoplus: improved protein function prediction from sequence // Bioinformatics. – 2020. – Vol. 36. – P. 422-429.

27 Голенко Е.С., Исмаилова А.А. Разработка R-скрипта для обработки данных масс-спектрометрии // Матер. международ. науч.-теорет. конф. «Сейфуллинские чтения – 16: Молодёжная наука новой формации – будущее Казахстана». – Нур-Султан. – 2020. – С. 149-151.

28 Голенко Е.С., Исмаилова А.А. Методы биоинформатики для подготовки и последующего анализа данных масс-спектрометрии // Матер. международ. науч.-практич. конф. «Интеграция науки, образования и производства основа реализации Плана Нации». – Караганда. – 2020. – С. 1011-1013.

29 Голенко Е.С., Исмаилова А.А. Метод вероятностного комбинирования результатов нескольких методологий поиска MC/MC для увеличения вероятности идентификации белков // Матер. международ. науч.-теорет. конф.

«Сейфуллинские чтения – 17: «Современная аграрная наука: цифровая трансформация»». – Нур-Султан. – 2021. – С. 215-218.

30 Голенко Е.С., Исмаилова А.А. Перспективы глубинного обучения в прогнозировании структуры белка // Матер. международ. науч. конф. «XXII Сатпаевские чтения». – Алматы. – 2022. – С. 495-502.

31 Голенко Е.С. Методы глубокого обучения для прогнозирования одномерных и двумерных аннотаций белка // Матер. международ. науч.-теорет. конф. «Сейфуллинские чтения – 18: «Молодёжь и наука – взгляд в будущее»». – Астана. – 2022. – С. 42-25.

32 Голенко Е.С., Исмаилова А.А. Модель машинного обучения для предсказания функций белков // Матер. международ. науч. конф. «Математическая логика и компьютерные науки». – Астана. – 2022. – С. 159-163.

33 Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank // Nucleic Acids Research. – 2013. – Vol. 41. – P. 36-42.

34 UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023 // Nucleic Acids Research. – 2023. – Vol. 51. – P. 523-531.

35 Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The Protein Data Bank // Nucleic Acids Research. – 2000. – Vol. 28. – P. 235-242.

36 Protein Data Bank Online Database // https://www.rcsb.org/. 01.12.2023.

37 Burgin J., Ahamed A., Cummins C., et al. The European Nucleotide Archive in 2022 // Nucleic Acids Research. – 2023. – Vol. 51. – P. 121-125.

38 Sonnhammer E.L., Eddy S.R., Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments // Proteins. – 1997. – Vol. 28. – P. 405-420.

39 Pfam // http://pfam.xfam.org/. 01.08.2023.

40 Zainab Noor et al. Mass spectrometry–based protein identification in proteomics – a review // Briefings in Bioinformatics. – 2021. – Vol. 22. P. 1620-1638.

41 Elias J.E., Gygi S.P. Target-decoy search strategy for mass spectrometry-based proteomics // Methods in Molecular Biology. – 2010. Vol. 604. – P. 55-71.

42 Spirin V., Shpunt A., Seebacher J., Gentzel M., Shevchenko A., Gygi S., Sunyaev S. Assigning spectrum-specific P-values to protein identifications by mass spectrometry // Bioinformatics. – 2011. – Vol. 27. – P. 1128-1134.

43 Голенко Е.С., Исмаилова А.А. Современные вычислительные стратегии для вывода белков в протеомике дробовика // Известия НАН РК. – № 336. – 2021. – С. 56-65.

44 Tang H., Arnold R.J., Alves P., Xun Z., Clemmer D.E., Novotny M.V., Reilly J.P., Radivojac P. A computational approach toward label-free protein quantification using predicted peptide detectability // Bioinformatics. - 2006. – Vol 22. – P. 481-488.

45 Nesvizhskii A.I., Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem // Molecular & Cellular Proteomics. – 2005. – Vol. 10. – P. 1419-1440.

46 Nesvizhskii A.I., Keller A., Kolker E., Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry // Analytical Chemistry. – 2003. – Vol. 75. – P. 4646-4658.

47 Golenko Y., Ismailova A., Rais Y. Protein Identification Using Sequence Databases // Scientific Journal of Astana IT University. – №4. – 2020. – C. 14-23.

48 Nesvizhskii A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics // Journal of Proteomics. – 2010. – Vol. 73. – P. 2092-2123.

49 NIST Libraries of Peptide Tandem Mass Spectra // https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:start. 01.01.2022.

50 Bjornson R.D., Carriero N.J., Colangelo C., Shifman M., Cheung K.H., Miller P.L., Williams K. X!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers // Journal of Proteome Research. – 2008. – Vol. 7. P. 293-299.

51 Fenyo D., Beavis R. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes // Analytical Chemistry. – 2003. – Vol. 75. – P. 768-774.

52 Phenyx Tools // https://www.phenyx-ms.com/.01.06.2023.

53 Colinge J., Masselot A., Giron M., Dessingy T., Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification // Proteomics. – 2003. – Vol. 3. – P. 1454-1463.

54 Gabaldon T., Huynen M.A. Prediction of protein function and pathways in the genome era // Cellular and Molecular Life Sciences. – 2004. – Vol. 61. – P. 930-944.

55 du Plessis L., Skunca N., Dessimoz C. The what, where, how and why of gene ontology - a primer for bioinformaticians // Briefings in Bioinformatics. – 2011. – Vol. 12. – P. 723-735.

56 Rost B., Liu J., Nair R., Wrzeszczynski K.O., Ofran Y. Automatic prediction of protein function // Cellular and Molecular Life Sciences. – 2003. – Vol. 60. – P. 2637-2650.

57 Shehu A., Barbara D., Molloy K. A survey of computational methods for protein function prediction // Big Data Analytics in Genomics. – 2016. – P. 225-298.

58 Bonetta R., Valentino G. Machine learning techniques for protein function prediction // Proteins: Structure, Function, and Bioinformatics. – 2020. – Vol. 88. – P. 397-413.

59 Zhao Y., Wang J., Chen J., Zhang X., Guo M., Yu G. A literature review of gene function prediction by modeling gene ontology // Frontiers in Genetics. – 2020. – Vol. 11.

60 Vu T.T.D., Jung J. Protein function prediction with gene ontology: from traditional to deep learning models // PeerJ. – 2020. – Vol. 9.

61 Golenko Y., Ismailova A., Shaushenova A., Mutalova Z., Dossalyanov D., Ainagulova A., Naizagarayeva, A. Implementation of machine learning models to determine the appropriate model for protein function prediction // Eastern-European Journal of Enterprise Technologies. – 2022. – Vol. 119. – P. 42-49.

62 Nariai N., Kolaczyk E.D., Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data // PLOS ONE. – 2007. – Vol. 3. – P. 337.

63 Kourmpetis Y.A., Van Dijk A.D., Bink M.C., Van Ham R.C., ter Braak CJ. Bayesian markov random field analysis for protein function prediction based on network data // PLOS ONE. – 2010. – Vol. 5. – P. 92-93.

64 Pinoli P., Chicco D., Masseroli M. Computational algorithms to predict gene ontology annotations // BMC Bioinformatics. – 2015. – Vol. 16. – P. 1-15.

65 Vinayagam A., del Val C., Schubert F., Eils R., Glatting K-H., Suhai S., König R. Gopet: a tool for automated predictions of gene ontology terms // BMC Bioinformatics. – 2006. – Vol. 7. – P. 161.

66 Lobley A.E., Nugent T., Orengo C.A., Jones D.T. Ffpred: an integrated featurebased function prediction server for vertebrate proteomes //Nucleic Acids Research. – 2008. – Vol. 36. – P. 297-302.

67 Lobley A., Swindells M.B., Orengo C.A., Jones D.T. Inferring function using patterns of native disorder in proteins // PLOS Computational Biology. – 2007. – Vol. 8. – P. 162.

68 Freund Y., Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. – 1997. – Vol. 55. – P. 119-139.

69 Toronen P., Medlar A., Holm L. Pannzer2: a rapid functional annotation web server // Nucleic Acids Research. – 2018. – Vol. 46. – P. 84-88.

70 You R., Zhang Z., Xiong Y., Sun F., Mamitsuka H., Zhu S. Golabeler: improving sequence-based large-scale protein function prediction by learning to rank // Bioinformatics. – 2018b. – Vol. 34. – P. 2465-2473.

71 Szklarczyk D., Morris J.H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N.T., Roth A., Bork P., Jensen L.J., Von Mering C. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible // Nucleic Acids Research. – 2016. – Vol. 45. – P. 362-368.

72 Голенко Е.С., Исмаилова А.А. Применение методов глубокого обучения для предсказания структуры белков // Вестник Национальной инженерной академии РК. – №4. – 2022. – С. 28-40.

73 Fa R., Cozzetto D., Wan C., Jones D.T. Predicting human protein function with multi-task deep neural networks // PLOS ONE. – 2018. – Vol. 13.

74 Cai Y., Wang J., Deng L. Sdn2go: an integrated deep learning model for protein function prediction // Frontiers in Bioengineering and Biotechnology. – 2020. – Vol. 8. – P. 391.

75 Du Z., He Y., Li J., Uversky V.N. Deepadd: protein function prediction from k-merembedding and additional features // Computational Biology and Chemistry. – 2020. – Vol. 89.

76 Голенко Е.С., Исмаилова А.А., Жумаханова А.С. Предсказание функций белков при помощи базы данных «Gene Ontology» и моделей машинного обучения // Известия НАН РК. – №2. – 2022. – С. 19-38.

77 Baldi P. Autoencoders, unsupervised learning, and deep architectures, In: Proceedings of ICML workshop on unsupervised and transfer learning // JMLR Workshop and Conference Proceedings. – 2012. – Vol. 2. – P. 37-49.

78 Chicco D., Sadowski P., Baldi P. Deep autoencoder neural networks for gene ontology annotation predictions // Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics. – 2014. – P. 533-540.

79 Gligorijevic V., Barot M., Bonneau R. Deepnf: deep network fusion for protein function prediction // Bioinformatics. – 2018. – Vol. 34. P. 3873-3881.

80 Peng J., Xue H., Wei Z., Tuncali I., Hao J., Shang X. Integrating multi-network topology for gene function prediction using deep neural networks // Briefings in Bioinformatics. – 2020. Vol. 22. P. 2096-2105

81 Zou X., Wang G., Yu G. Protein function prediction using deep restricted Boltzmann machines // BioMed Research International. – 2017.

82 Seyyedsalehi S.F., Soleymani M., Rabiee H.R., Mofrad M.R. Pfp-wgan: protein function prediction by discovering gene ontology term correlations with generative adversarial networks // PLOS ONE. – 2021. – Vol. 16.

83 Cao R., Freitas C., Chan L., Sun M., Jiang H., Chen Z. Prolango: protein function prediction using neural machine translation based on a recurrent neural network // Molecules. – 2017. – Vol. 22. – P. 1732.

84 Nauman M., Rehman H.U., Politano G., Benso A. Beyond homology transfer: deep learning for automated annotation of proteins // Journal of Grid Computing. – 2019. – Vol. 17. – P. 225-237.

85 Seyyedsalehi S.F., Soleymani M., Rabiee H.R., Mofrad M.R. Pfp-wgan: protein function prediction by discovering gene ontology term correlations with generative adversarial networks // PLOS ONE. – 2021. – Vol. 16.

86 Tavanaei A., Maida A.S., Kaniymattam A., Loganantharaj R. Towards recognition of protein function based on its structure using deep convolutional networks // 2016 IEEE international conference on bioinformatics and biomedicine. – 2016. – P. 145-149

87 Kulmanov M., Khan M.A., Hoehndorf R. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier // Bioinformatics. – 2018. – Vol. 34. – P. 660-668.

88 Li J., Wang L., Zhang X., Liu B., Wang Y. Gonet: a deep network to annotate proteins via recurrent convolution networks // 2020 IEEE international conference on bioinformatics and biomedicine. – 2020. – Vol. 2. – P. 29-34.

89 Zhang F., Song H., Zeng M., Li Y., Kurgan L., Li M. Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions // Proteomics. – 2019. – Vol. 19.

90 Wan C., Jones D.T. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks // Nature Machine Intelligence. – 2020. – Vol. 9. – P. 540-550.

91 Tiwary S., Levy R., Gutenbrunner P., Soto F.S., Palaniappan K.K., Deming L., et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis // Nature methods. – 2019. – Vol. 16. – P. 519-525.

92 Gessulat S., Schmidt T., Zolg D.P., Samaras P., Schnatbaum K., Zerweck J., et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning // Nature methods. – 2019. – Vol. 16. – P. 509-518.

93 Zhou X.X., Zeng W.F., Chi H., Luo C., Liu C., Zhan J., et al. pdeep: Predicting MS/MS spectra of peptides with deep learning // Analytical chemistry. – 2017. – Vol. 89. – P. 12690-12697.

94 Diament B.J., Noble W.S. Faster SEQUEST searching for peptide identification from tandem mass spectra // Journal of proteome research. – 2011. – Vol. 10. – P. 3871-3879.

95 Gabriels R., Martens L., Degroeve S. Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques // bioRxiv. – 2019. – P. 295-299.

96 Kong A.T., Leprevost F.V., Avtonomov D.M., Mellacheruvu D., Nesvizhskii A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics // Nature methods. – 2017. – Vol. 14. – P. 513-520.

97 Tariq M.U., Saeed F. SpeCollate: Deep cross-modal similarity network for mass spectrometry data-based peptide deductions // PLOS ONE. – 2021.

98 Faghri F., Fleet D.J., Kiros J.R., Fidler S. Vse++: Improving visual-semantic embeddings with hard negatives // arXiv preprint arXiv:170705612. – 2017.

99 Schroff F., Kalenichenko D., Philbin J. Facenet: A unified embedding for face recognition and clustering // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – P. 815-823.

100 Wang L., Li Y., Huang J., Lazebnik S. Learning two-branch neural networks for image-text matching tasks // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2018. – Vol. 41. – P. 394-407.

101 Nam H., Ha J.W., Kim J. Dual attention networks for multimodal reasoning and matching // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – P. 299-307.

102 Qin C., Luo X., Deng C., Shu K., Zhu W., Griss J., et al. Deep learning embedder method and tool for mass spectra similarity search // Journal of Proteomics. – 2021. – P. 232

103 May D.H., Bilmes J., Noble W.S. A learned embedding for efficient joint analysis of millions of mass spectra // BioRxiv. – 2018. – P. 675-678.

104 Schultz M., Joachims T. Learning a distance metric from relative comparisons // Advances in neural information processing systems. – 2004. – P. 41-48.

105 Schroff F., Kalenichenko D., Philbin J. Facenet: A unified embedding for face recognition and clustering // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – P. 815-823.

106 Sharma K., D'Souza R.C., Tyanova S., Schaab C., Wiśniewski J.R., Cox J., et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling // Cell reports. – 2014. – Vol. 8. – P. 1583-1594.

107 Bittremieux W., Meysman P., Noble W.S., Laukens K. Fast open modification spectral library searching through approximate nearest neighbor indexing // Journal of proteome research. – 2018. – Vol. 17. – P. 3463-3474.

108 Chick J.M., Kolippakkam D., Nusinow D.P., Zhai B., Rad R., Huttlin E.L., et al. A mass-tolerant database search identifies a large proportion of unassigned

spectra in shotgun proteomics as modified peptides // Nature biotechnology. – 2015. – Vol. 33. – P. 743-749.

109 Park C.Y., Klammer A.A., Kall L., MacCoss M.J., Noble W.S. Rapid and accurate peptide identification from tandem mass spectra // Journal of proteome research. – 2008. – Vol. 7. – P. 3022-3027.

110 Kall L., Canterbury J.D., Weston J., Noble W.S., MacCoss M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets // Nature methods. – 2007. – Vol. 11. – P. 923-925.

111 Bileschi M.L., Belanger, D., Bryant D., Sanderson T., Carter B., Sculley D., DePristo M.A., Colwell L.J. Using Deep Learning to Annotate the Protein Universe. // bioRxiv. – 2019.

112 Голенко Е.С., Исмаилова А.А. Предсказание функций белка с использованием комбинации BiLSTM и алгоритма самовнимания // Известия НАН РК. – №3. – 2023. – С. 62-75.

113 Becker J., Maes F., Wehenkel L. On the Encoding of Proteins for Disordered Regions Prediction // PlOS ONE. – 2013.

114 IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents // Biochemistry. - 1970. – P. 4022-4027.

115 Swiss-Prot Database // https://www.uniprot.org/. 01.12.2023.

116 Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. – 2005. – Vol. 18. – P. 602-610.

117 Abduljabbar R.L., Dia H., Tsai P. Unidirectional and Bidirectional LSTM Models for Short-Term Traffic Prediction // Journal of Advanced Transportation. – 2021. – Vol. 4. – P. 1-16.

118 Kurtukova A.V., Romanov A.S. Modeling the neural network architecture to identify the author of the source code // Proceedings of Tomsk State University of Control Systems and Radioelectronics. – 2019. – Vol. 22. – P. 37-42.

119 Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. – 2017. – P. 6000-6010.

120 Mnih V., Kavukcuoglu K., et al. Recurrent Models of Visual Attention // NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems. – 2014. – Vol. 2. – P. 2204-2212.

121 Yang Z. et al. Hierarchical Attention Networks for Document Classification // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2016. – P. 1480-1489.

122 Bahdanau D. et al. Neural Machine Translation by Jointly Learning to Align and Translate // ArXiv. 1409. – 2014.

123 Verga et al. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction // NAACL 2018. – 2018.

ПРИЛОЖЕНИЕ А

Авторские свидетельства

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 42935 от «15» февраля 2024 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
ГОЛЕНКО ЕКАТЕРИНА СЕРГЕЕВНА

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Программный модуль «PFP_SelfAttn_BiLSTM»**

Дата создания объекта: **14.02.2024**

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

Подписано ЭЦП

Е. Оспанов

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 44224 от «3» апреля 2024 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
ГОЛЕНКО ЕКАТЕРИНА СЕРГЕЕВНА, ИСМАИЛОВА АЙСУЛУ АБЖАППАРОВНА

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Программный модуль «SC_MS_Peptide_Ident»**

Дата создания объекта: **01.04.2024**

РГП «Национальный институт интеллектуальной собственности» МЮ РК

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

Подписано ЭЦП

Е. Оспанов

ПРИЛОЖЕНИЕ Б

Акты внедрения

«ҰЛТТЫҚ БИОТЕХНОЛОГИЯ ОРТАЛЫҒЫ» ЖАУАПКЕРШІЛІГІ ШЕКТЕУЛІ СЕРІКТЕСТІГІ
010000, Астана қ., Есіл ауданы, Қорғалжын тас жолы, 13/5
телефон: +7(7172) 70-75-65
e-mail: info@biocenter.kz

ТОВАРИЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ «НАЦИОНАЛЬНЫЙ ЦЕНТР БИОТЕХНОЛОГИИ»
010000, г. Астана, Есильский район, Шоссе Коргалжын, 13/5
телефон: +7(7172) 70-75-65
e-mail: info@biocenter.kz

№ 12-39-119
«10» 04 2024 ж.

СПРАВКА О ВНЕДРЕНИИ
результатов исследовательской работы в научную деятельность

Соискателем ученой степени доктора философии (PhD), м.т.н. Голенко Е.С. в рамках выполнения диссертационной работы предложен алгоритм для идентификации пептидов на основании их масс-спектров и разработан соответствующий программный модуль «SC_MS_Peptide_Ident».

В результате проведенных исследований была предложена модель машинного обучения, принимающая на входе два набора данных, состоящие из масс-спектров и пептидов, проведено обучение нейронной сети и оценка показателей, разработан программный модуль «SC_MS_Peptide_Ident», позволяющий как обучать модель на предложенных данных, так и использовать уже обученную нейронную сеть для сопоставления масс-спектров пептидов с их аминокислотными последовательностями.

Отработанные соискателем Голенко Е.С. методы исследований и полученные научные результаты отражены в материалах международных научно-практических конференций, научных статьях отечественных журналов рекомендованных КОКСНВО и зарубежных изданиях, входящих в базу данных Scopus.

Разработанные программные модули для идентификации пептидов и функций белков внедрены в научный процесс лаборатории биоразнообразия и генетических ресурсов Национального центра биотехнологии.

И.о. генерального директора,
к.б.н., профессор **Огай В.Б.**

Заведующий лабораторией биоразнообразия
и генетических ресурсов,
PhD, ассоциированный профессор **Киян В.С.**

000072

ООО «Новые Программные Системы»
630090, г. Новосибирск, пр-т Ак.Лаврентьева, д.6, оф. 222 Тел. +7(383)332 1676
ОГРН 1075473011900 ИНН/КПП 5408254508/540801001 ОКПО 82284174
р/с № 40702810695240000432 в Филиал Сибирский ПАО Банк «ФК Открытие», БИК 045004867

АКТ
внедрения результатов научно-исследовательской деятельности
PhD докторанта по направлению подготовки кадров
«8D061 – Информационно-коммуникационные технологии»
Образовательной программы «Аналитика больших данных»
Казахского агротехнического исследовательского университета имени С. Сейфуллина
ГОЛЕНКО ЕКАТЕРИНЫ СЕРГЕЕВНЫ

Мы, нижеподписавшиеся, биоинформатик Вяткин Юрий Викторович, программист Голосов Кирилл Владимирович, составили настоящий акт о том, что результаты диссертационной работы Голенко Е. С. «Разработка алгоритмов анализа данных масс-спектрометрии нативных белков», а именно разработанный в рамках исследования программный модуль PFP_SelfAttn_BiLSTM, прошёл производственную проверку и внедрён в процесс исследований в компании ООО «Новые Программные Системы».

Программный модуль служит инструментом дополнительной проверки результатов экспериментов и исследований, проводимых по функциональному аннотированию.

Председатель комиссии
Мигинский Денис Сергеевич _____ Заместитель директора по науке

Члены комиссии:

Вяткин Ю. В. _____ Биоинформатик

Голосов К. В. _____ Программист

ООО «Новые Программные Системы»
630090, г. Новосибирск, пр-т Ак.Лаврентьева, д.6, оф. 222 Тел. +7(383)332 1676
ОГРН 1075473011900 ИНН/КПП 5408254508/540801001 ОКПО 82284174
р/с № 40702810695240000432 в Филиал Сибирский ПАО Банк «ФК Открытие», БИК 045004867

АКТ
внедрения результатов научно-исследовательской деятельности
PhD докторанта по направлению подготовки кадров
«8D061 – Информационно-коммуникационные технологии»
Образовательной программы «Аналитика больших данных»
Казахского агротехнического исследовательского университета имени С. Сейфуллина
ГОЛЕНКО ЕКАТЕРИНЫ СЕРГЕЕВНЫ

Мы, нижеподписавшиеся, биоинформатик Вяткин Юрий Викторович, программист Голосов Кирилл Владимирович, составили настоящий акт о том, что результаты диссертационной работы Голенко Е. С. «Разработка алгоритмов анализа данных масс-спектрометрии нативных белков», а именно разработанный в рамках исследования программный модуль SC_MS_Peptide_Ident, прошёл производственную проверку и внедрён в процесс исследований в компании ООО «Новые Программные Системы».

Программный модуль служит инструментом дополнительной проверки результатов экспериментов и исследований, проводимых по идентификации пептидов из данных масс-спектрометрии.

Председатель комиссии
Мигинский Денис Сергеевич _____ Заместитель директора по науке

Члены комиссии:

Вяткин Ю. В. _____ Биоинформатик

Голосов К. В. _____ Программист