Цели и задачи ПО

Программа предназначена для принятия решений в области предсказания геномных функциональных элементов (ГФЭ) с использованием нейросетевых моделей глубинного обучения разных типов архитектур в целях применения пользователем, не обладающим глубокими навыками программирования и/или не имеющего опыта применения алгоритмов искусственного интеллекта.

Программа реализует предсказание геномной разметки анализируемого ГФЭ на основе экспериментальных данных расположения этого ГФЭ в отдельных участках генома, а также других омиксных данных, подгружаемых пользователем с применением моделей на основе искусственного интеллекта.

В программе интегрированы пять функциональных модулей, каждый из которых разработан для решения отдельного класса задач, определяемых в названии модуля.

- Модуль №1. «Предсказание ГФЭ на основе омиксных данных с использованием нейронных сетей разных типов, а именно сверточных, рекуррентных, гибридных сверточно-рекуррентных, генеративно-состязательных, графовых, а также нейронных сетей типа трансформер»
- Модуль №2. «Формирование и выборочный отбор омиксных данных с целью создания входных данных для моделей глубинного обучения»
- Модуль №3. «Анализ на ассоциацию геномных признаков с ГФЭ»
- Модуль №4. «Трансферное обучение для переноса аннотации ГФЭ»
- Модуль №5. «Интерпретация обученных нейросетевых моделей»

Целью создания единого программного обеспечения, объединяющего программные модули, являются обеспечение возможности выбора лучшего предсказания среди предсказаний моделей, разработанных на основе нейронных сетей разных типов и обученных на данных пользователя, и возможности настройки моделей в соответствии с желаемыми результатами, а

также возможность дообучения индивидуальных моделей и предоставления возможности их использования другими пользователями.

Затрачиваемые ресурсы для работы

Затрачиваемые ресурсы определяются минимальными требования для случая разворачивания ПО на одном сервере:

- Количество физических ядер процессора (CPU) - 4;
- Оперативная память (RAM) - 32 Гб;
- Дисковое пространство (HDD) – 256 Гб;
- GPU класса Nvidia, GTX-1080Ti;
- Сервер должен быть подключен к ИБП
- Операционная система Ubuntu версии 22.XX и выше.

Вводная информация, входные и выходные данные

Модуль №1 «Предсказание ГФЭ с использованием нейронных сетей разных типов, а именно сверточных, рекуррентных, гибридных сверточно-рекуррентных, генеративно-состязательных, графовых, а также нейронных сетей типа трансформер»:

Модуль предоставляет функционал по предсказанию вероятности нахождения ГФЭ для каждой позиции или области генома с использованием сверточных; рекуррентных; гибридных сверточно-рекуррентных, включающих комбинацию из сверточных и рекуррентных слоев; генеративно-состязательных или графовых нейронных сетей, а также нейронных сетей типа трансформер, реализованных с использованием двух принципиально разных моделей.

Входные данные:

- Загруженные пользователем файлы геномной аннотации в формате BED, загруженные геномные сборки в формате FASTA и необходимые омиксные данные в формате BED

Выходные данные:

- Файл в формате BigWig, где для каждой позиции генома выводится вероятность нахождения функционального элемента в данной позиции
- Файл с предсказаниями в формате BED

Модуль №2 «Формирование и выборочный отбор омиксных данных с целью создания входных данных для моделей глубинного обучения»:

Модуль представляет собой алгоритмическое решение для формирования и выборочного отбора омиксных данных для последующей обработки методами глубинного обучения. Библиотека реализует выборочный отбор омиксных данных, основанный на заданных пользователем параметрах выборки, а также формирует и подготавливает данные для дальнейшей обработки методами глубинного обучения.

Входные данные:

- Источник омиксных данных, набор параметров, определяющих критерии отбора омиксных данных

Выходные данные:

Архив «tar.gz», в котором содержатся файлы в формате BED с экспериментами, удовлетворяющие заданным параметрам

Модуль №3 «Анализ на ассоциацию геномных признаков с ГФЭ»:

Модуль предоставляет функционал по исследованию геномных признаков на ассоциацию.

Входные данные:

- Исследуемые геномные признаки в формате BED
- Файл chrom.sizes с размерами хромосом для исследуемого генома

Выходные данные:

- Таблица с результатами статистического анализа

Модуль №4 «Трансферное обучение для переноса аннотации ГФЭ»:

Модуль предлагает пайплайн, предоставляющий функционал для генерации межвидовой аннотации функциональных элементов генома (вторичных структур ДНК, гистоновых меток, транскрипционных факторов и т.д.). Результатом работы модуля является сгенерированная аннотация ГФЭ для организма, не имеющего на данный момент экспериментально подтвержденной аннотации данного элемента.

Входные данные:

- Аннотация функционального элемента в исходном геноме в формате BED
- Полная последовательность генома, для которого требуется перенос аннотации

Выходные данные:

- Аннотация участков в заданном геноме в формате BED

Модуль №5 «Интерпретация обученных нейросетевых моделей»:

Модуль предлагает гибкий пайплайн, интерпретирующий обученные модели глубинного обучения для предсказания расположения ГФЭ, разработанные на основе нейронных сетей разных типов.

Входные данные:

- Модель глубинного обучения, обученная для предсказания расположения ГФЭ
- Тензор, для каждой позиции генома выводится вероятность нахождения функционального элемента в данной позиции

Выходные данные:

- Карта значимости омиксных данных, которая представляет из себя тензор уровня вклада каждого признака для каждой позиции генома. После усреднения по цепочке и по всей выборке можно получить усредненный уровень значимости в определении ГФЭ для каждого признака из омиксных данных