- **Тема работы**
  Разработка модели для предсказания пространственных контактов хроматина на основе эпигенетических характеристик геномов мыши и человека на основе методов машинного обучения

- **Состав коллектива:**
  Фишман Вениамин Семенович, ИЦиГ СО РАН, НГУ, кбн
  Белокопытова Полина Станиславовна, ИЦиГ СО РАН, НГУ
  Нуриддинов Мирослав Абдурахимович, ИЦиГ СО РАН
  Можейко Евгений Александрович, ИЦиГ СО РАН

- **Научное содержание работы:**

1. *Постановка задачи.*

Только 2% от всей длины ДНК занимают белок кодирующие последовательности, оставшаяся часть молекулы представлена некодирующими элементами, в число которых входят такие регуляторные участки как энхансеры и промоторы. Активность энхансеров является клеточно-специфичной и способна изменять уровень экспрессии генов в сотни и тысячи раз.

Специфичность действия энхансеров определяется физическим взаимодействием с промоторной областью гена, поэтому для реализации функции необходимо, чтобы энхансеры и регулируемые ими промоторы были сближены в пространстве ядра. Однако энхансеры часто располагаются на огромном расстоянии (до 1 МБ) от промоторов, если рассматривать их в линейных координатах генома. Предполагается, что компактно уложенный геном имеет петлевую организацию, таким образом отдаленные регионы оказываются в пространстве рядом друг с другом. В ряде работ показано, что наблюдаемые при хромосомных перестройках изменения пространственных контактов между регуляторными элементами могут приводить к развитию заболеваний, даже если сами регуляторные элементы и кодирующие участки генома сохраняются неизменными, при чем патологический фенотип, сопровождающий хромосомную перестройку, является следствием нарушений промотор-энханесрных взаимодействий. В связи с этим исследование нарушений трехмерной организации в перестроенных геномах представляет значительный интерес.

Существуют методы, которые позволяют экспериментально исследовать трехмерную организацию генома. Это такие методы как Hi-C, CHIA-PET и другие технологии, основанные на 3C. Однако эти исследования являются достаточно дорогими, поэтому для менее исследованных типов клеток часто таких данных нет, зато доступны различные эпигенетические данные. Поэтому важно изучить какие закономерности лежат в основе формирования пространственной организации генома и научиться предсказывать трехмерную архитектуру на основе эпигенетических данных, особенно это актуально в случае хромосомных перестроек.

Целью нашей работы является разработка алгоритма для предсказания трехмерных контактов хроматина в нормальных клетках и клетках с хромосомными перестройками на основе доступных эпигенетических данных. Для этого мы используем методы, основанные на алгоритмах машинного обучения.

2. *Современное состояние проблемы (на момент начала работы).*

Тема предсказания трехмерной архитектуры генома на основе эпигенетических, физических и нуклеотидных характеристик генома сейчас является очень актуальной. Подходы к автоматизированному анализу последствий хромосомных перестроек в контексте трехмерной архитектуры ядра начали развиваться в 2016 году. Был предложен алгоритм, предсказывающий, какие хромосомные перестройки ассоциированы с болезнями, вызванными изменением экспрессии генов в результате изменения 3D структуры хроматина. Исследователи использовали оценку SI (Structure Influence), которая количественно определяет степень влияния хромосомной перестройки на пространственную структуру хроматина. Нужно отметить, что использованный авторами алгоритм предполагает знание о расположении ТАД в геноме. Технологии, основанные на 3C, такие как Hi-C и CHIA-PET, дают подробную информацию о пространственной структуре хроматина, однако эти технологии являются достаточно дорогостоящими. Поэтому разрабатывают новые технологии, которые предсказывают пространственную организацию генома, а также взаимодействие основных регуляторных элементов используя эпигенетические метки и машинное обучение. Внутри клеточного ядра геном складывается в организованные структуры, характерные для каждого клеточного типа. Было показано, что эта архитектура хроматина может быть предсказана de novo, используя эпигенетические данные, полученные с помощью Chip-seq. В основе этого исследования лежит идея, что хромосомы кодируют 1D-последовательность структурных типов хроматина. Взаимодействия между этими типами хроматина определяют трехмерный структурный ансамбль посредством процесса, аналогичного разделению различных фазовых (агрегатных) состояний вещества. Нейронная сеть используется для определения связи между эпигенетическими метками, присутствующими в локусе, и компартментом к которому относится данный локус.

Как известно, CTCF играет значительную в формировании петель и пространственной архитектуры генома. Был предложен алгоритм машинного обучения, предсказывающий петли, опосредованные CTCF. В качестве признаков использовались данные по экспрессии, эпигененические модификации гистонов, Chip-seq данные архитектурных белков, данные DNAse-seq, длина петель, ориентация CTCF сайтов и т.д. Также, сейчас популярны работы, предсказывающие пространственную организацию генома в общем в виде карт контактов, в результате чего можно предсказывать последствия хромосомных перестроек, основываясь на изменении взаимодействий между конкретным энхансером и промотором. Есть несколько подходов для достижения этой цели. Например, в одной из работ авторы предсказывают трехмерную карту контактов, используя построенную физическую модель, описывающую трехмерные контакты генома. В результате с высокой точностью можно предсказать 3D организацию генома и последствия хромосомных перестроек, однако для построения физической модели необходимы данные Hi-C, которые есть далеко не для каждого типа клеток.Помимо физического моделирования часто используют методы машинного обучения для предсказания трехмерной карты контактов. В одной из работ с помощью машинного обучения предсказывались взаимодействия промоторов и энхансеров, используя только лишь последовательность нуклеотидов. Последовательности нуклеотидов фиксированной длины представлялись в виде векторов и обрабатывались при помощи методов обучения, используемых в такой области искусственного интеллекта как обработка естественного языка. После этого с помощью классификаторов обучали на взаимодействующих и не взаимодействующих промотор-энхансерных парах [27]. В момент выполнения нашей

работы, одним из наиболее эффективных алгоритмов для предсказания трехмерных взаимодействий промоторов и энхансеров являлся TargetFinder. Алгоритм TargetFinder, предсказывающий взаимодействие пары промотор-энхансер на основании Chip-seq данных различных белков, DNAse-seq, транскрипцион-ных данных и т.д

Несмотря на успех нейронных сетей и машинного обучения, остается не понятным что является причиной, а что следствием. Или эпигенетические метки расположены в данных участках генома из-за принятой конформации хроматина, или пространственная архитектура такая из-за присутствия этих белков и эпигенетических меток. Также остается вопросом можно ли предсказать, основываясь на Chip-seq данных, что произойдет с экспрессией генов при какой либо хромосомной перестройке.

3. *Подробное описание работы, включая используемые алгоритмы.*

   В работе использовались алгоритмы машинного обучения из библиотеки XGBoost. Все скрипты для генерации обучающей и тестовой выборки были написаны на языке Python. Эпигенетические данныеb Hi-C были скачаны из доступных баз данных (NCBI, ENCODE).

4. *Полученные результаты.*
   - На основе методов машинного обучения нами разработан алгоритм, позволяющий с высокой точностью предсказывать трехмерные контакты хроматина, включая промотор-энхансерные взаимодействия в клетках мыши и человека, на основе данных о генной экспрессии, распределении и ориентации сайтов связывания белка CTCF и геномных расстояний между контактирующими участками (Рис.1).
   - Использование разработанного алгоритма для моделирования трехмерных контактов хроматина в клетках почки конечности с делецией в локусе EPHA4 показало, что предсказанные частоты контактов соответствуют экспериментальным данным об эктопических взаимодействиях, формирующихся вследствие делеции (Рис.2).

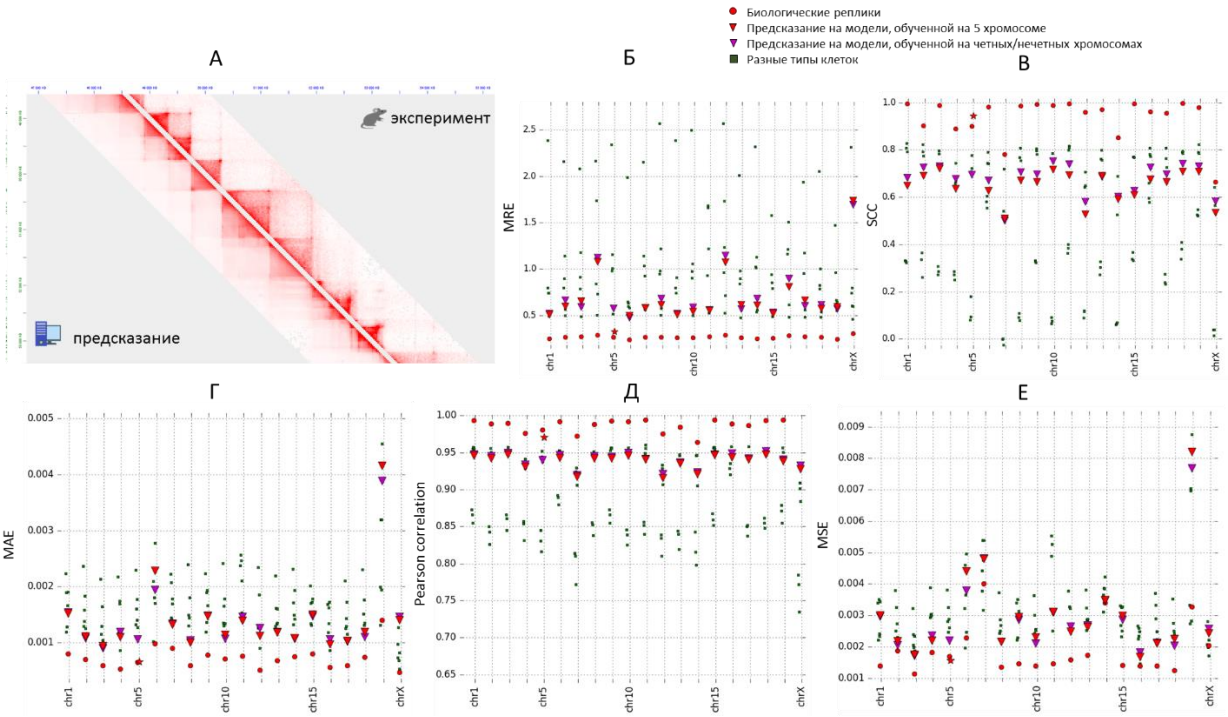

Рис.1. Высокая точность предсказания пространственной организации генома для гепатоцитов мыши. (А) Модель обучена на четных и нечетных хромосомах, выборка для предсказания не пересекается с обучающей выборкой. Карта контактов предсказанная (снизу) и полученная экспериментально (сверху) для 2 хромосомы. (Б, В, Г, Д, Е) Метрики, полученные при сравнении предсказанных контактов гепатоцитов мыши с экспериментальными данными.

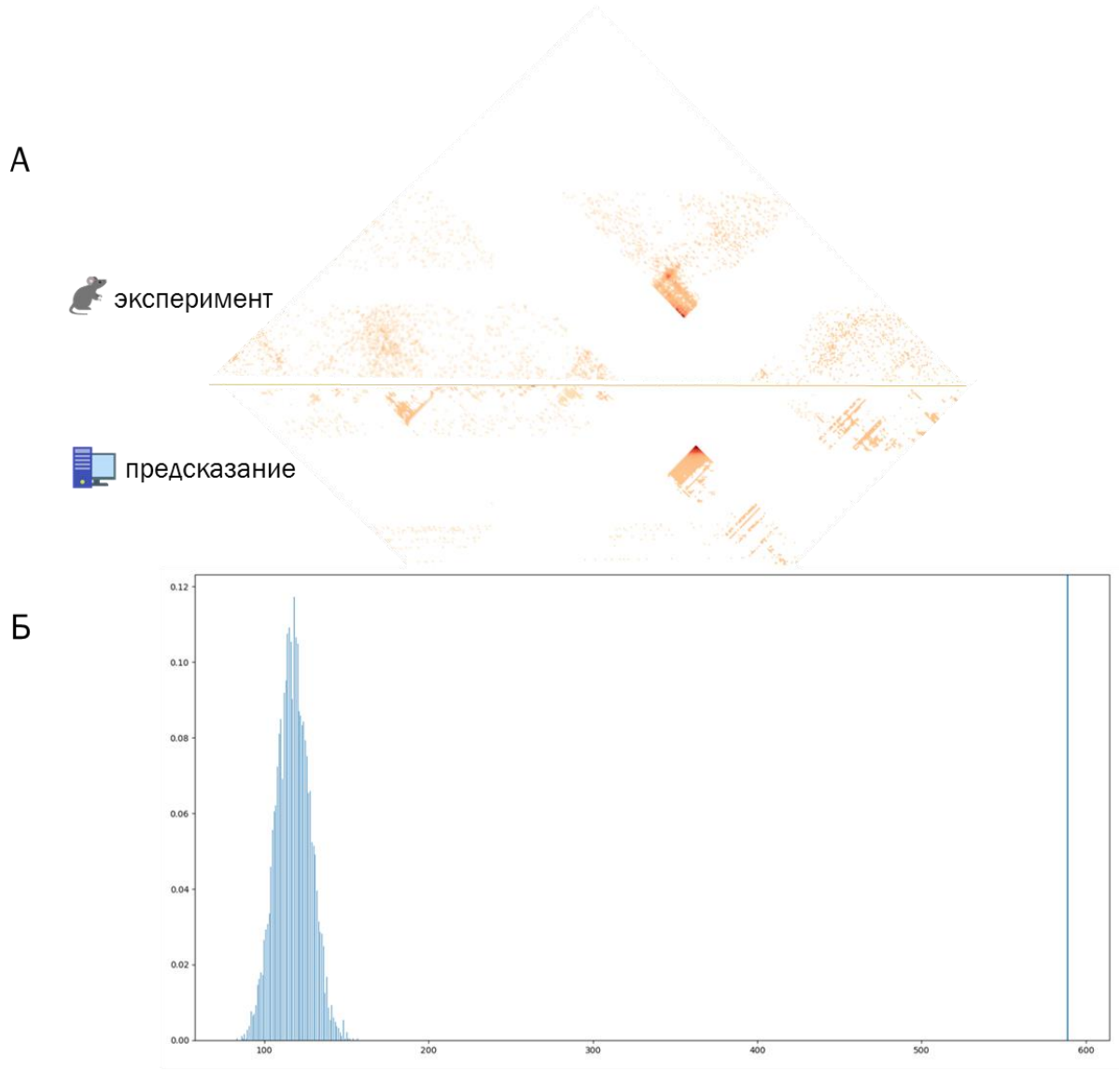

Рис. 2. Разработанный алгоритм позволяет предсказывать изменения в пространственной организации генома при перестройке. (А) Эктопические взаимодействия для экспериментальных данных и предсказанных нашим алгоритмом. (Б) Распределение случайных контактов, пересекающихся с реальными эктопическими, чертой отмечено количество пересечений предсказанных эктопических контактов с реальными эктопическими.

- **Эффект от использования кластера в достижении целей работы.**

Многие биоинформационные программы требуют много оперативной памяти, также как и некоторые вычисления точности предсказаний, используемые в разработанном алгоритме. Без кластера подобные вычисления были бы невозможны.