# CHiMAERa: инструмент для интерпретируемого предсказания карт Hi-C

Школиков А.С.

*ФББ МГУ, Москва, Ленинские горы, 1с73,* a.shkolikov@gmail.com

Гельфанд М.С.

*ИППИ РАН, Москва, Большой Каретный пер., 19с1,* mikhail.gelfand@gmail.com

Хроматин в интерфазном ядре имеет довольно сложную организацию. Изучение механизмов укладки хроматина в ядре может пролить свет на многие вопросы молекулярной биологии, связанные с реализацией генетической информации в клетке. Мы создали инструмент для предсказания контактных карт Hi-C по последовательности ДНК, названный нами CHiMAERa (**C**onvolutional neural net for **Hi**-C **m**aps prediction using **a**uto**e**ncoder for maps **r**epresentation). Инструмент основан на модели глубокого обучения, схема которой показана на рис. 1. С помощью данного инструмента можно не только получать предсказания, но и интерпретировать их, тем самым раскрывая, какие паттерны в нуклеотидной последовательности важны для формирования пространственной структуры хроматина. Эти данные, в свою очередь, могут помочь раскрыть биологические механизмы лежащие в основе закономерностей, выявленных моделью машинного обучения.

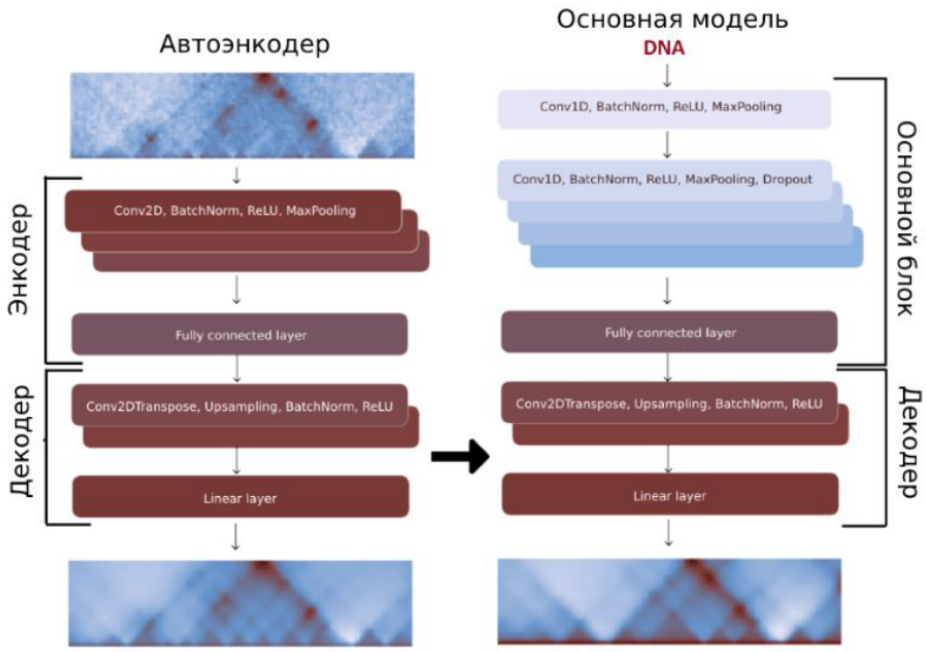

Рис. 1. Архитектура модели и стратегия обучения. Модель состоит из автоэнкодера, осуществляющего предобработку карт Hi-C и получающего их скрытые представления (слева), и основной модели, предсказывающей карты по последовательности ДНК. Основная модель на старте обучения уже содержит обученный декодер из автоэкодера, таким образом ее задача сводится к предсказанию скрытых представлений карт.

Хотя уже существуют модели машинного обучения, решающие задачу предсказания карт по последовательности[1,2,3], все они применялись для млекопитающих, а основной целью работ было само предсказание, в то время как интерпретации уделялось мало внимания. Наш инструмент на данный момент содержит модели, обученные на данных для 4 организмов - *Homo sapiens* (клеточная линия HFF), *Drosophila melanogaster, Saccharomyces cerevisiae* и *Dictyostelium discoideum.* Также инструмент позволяет проводить обучение на новых данных и содержит набор функций для их предобработки. Для всех моделей корреляция Пирсона между предсказанными и истинными картами на тестовых выборках составила более 0.65 (до 0.8 у *S. cerevisiae*); качество предсказаний для человека близко к полученному в работах, упомянутых выше (примеры предсказаний показаны на рис. 2).

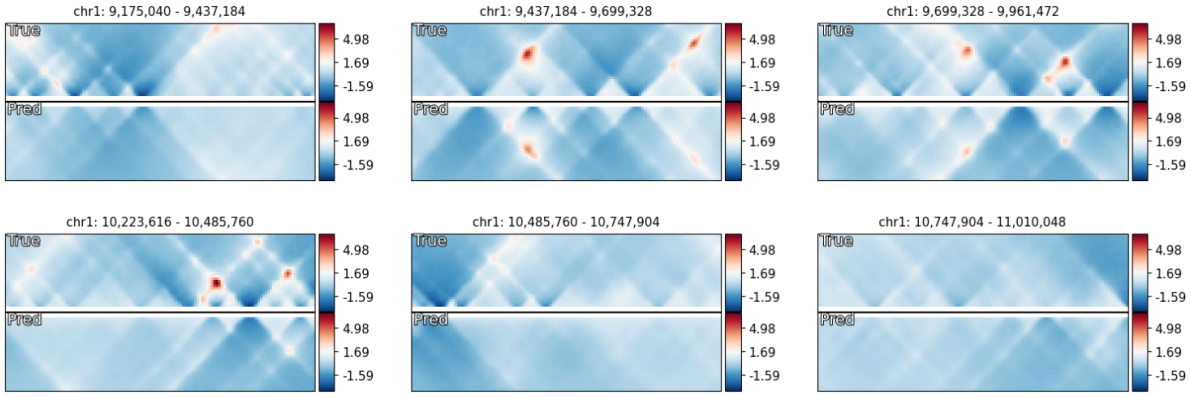


Рис. 2. Пример предсказаний для человека - верхняя половина каждого изображения содержит истинную карту, обработанную автоэнкодером, нижняя - предсказанную по нуклеотидной последовательности (для удобства показана зеркально)

Отличительной особенностью нашей модели является также использование автоэнкодера для предобработки карт Hi-C и получения их скрытого представления. В скрытом представлении могут быть получены вектора, соответствующие тем или иным структурам, представленным в картах. С помощью этих векторов может проводиться

поиск паттернов в нуклеотидной последовательности, ассоциированных с данными структурами.

Набор методов интерпретации предсказаний, включающий различные применения *in silico* мутагенеза, стохастические алгоритмы поиска мотивов, анализ градиентов модели, позволяет найти нуклеотидные паттерны, наличие которых в последовательности приводит к предсказанию моделью определенных структур в картах Hi-C. Для найденных мотивов может быть измерена их важность для предсказаний в сравнении с различными контролями. Также предусмотрен поиск закономерностей, не предусматривающих наличия коротких паттернов - например, основанных на взаимном расположении генов.

Для *H. sapiens* нами была показана важность сайтов CTCF (продемонстрировано на рис. 3) и YY1; для D. *melanogaster* - BEAF-32, m1bp и других, в том числе не известных в этой роли; для *S. cerevisiae* - reb1p. Для *D. discoideum* мотивов не обнаружено, но показана важность взаимного расположения генов (рис. 4).

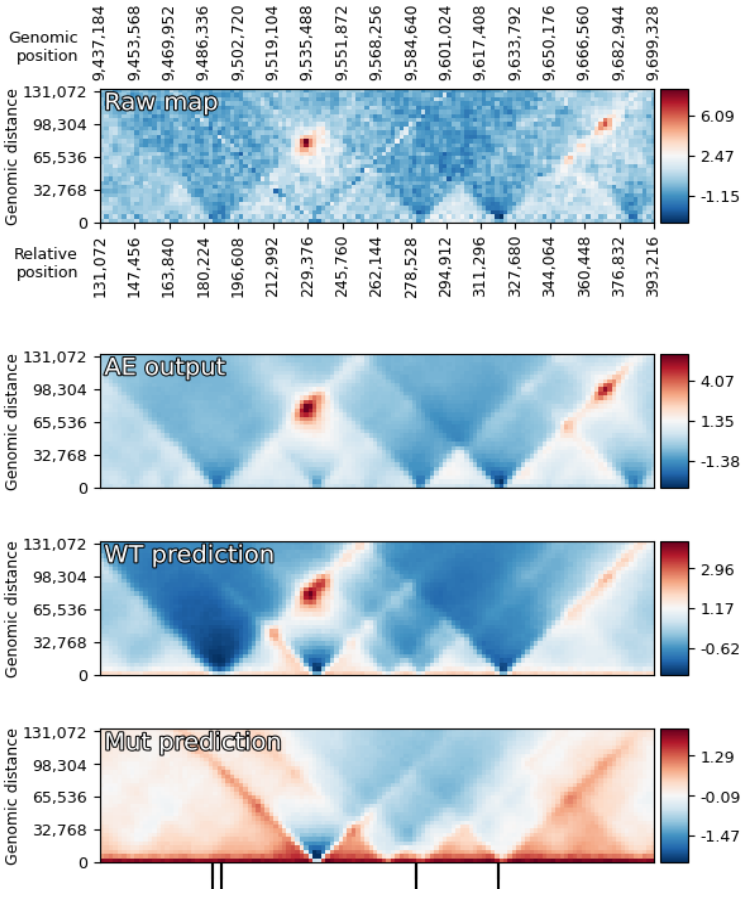

Рис. 3. Пример *in silico* мутагенеза на данных человека. Сверху вниз изображены: исходная карта; карта, обработанная автоэнкодером; карта, предсказанная по последовательности дикого типа; карта, предсказанная по последовательности, из которой удалены 4 сайта CTCF (позиции указаны стрелками), найденные с помощью метода интегрированных градиентов.

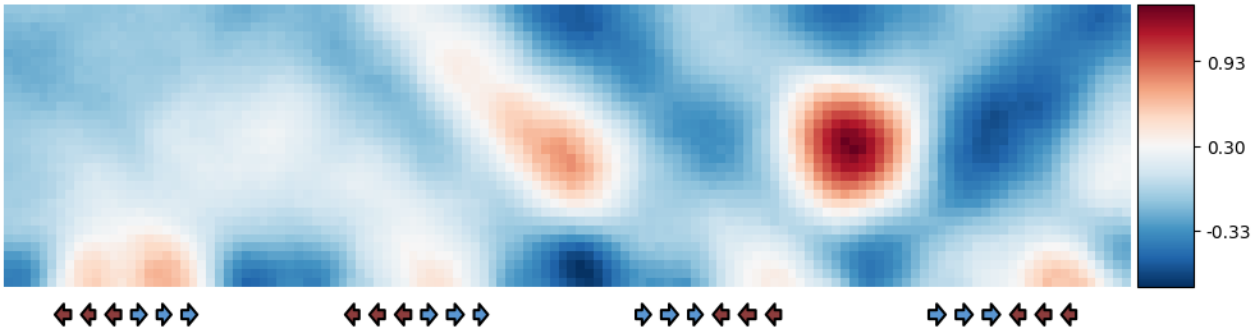


Рис. 4. Влияние взаимного расположения генов на предсказания у *D. discoideum*. Показано усредненное предсказание по химерным последовательностям, собранным из случайной выборки генов (показаны стрелками в зависимости от их направления) и межгенных участков (все вокруг стрелок). Между двумя участками конвергентно расположенных генов предсказывается повышенная частота контактов, тогда как для дивергентно расположенных генов этого не наблюдается.

1. G.Fudenberg, D.R.Kelley, K.S.Pollard (2020). Predicting 3D genome folding from DNA sequence with Akita, *Nature methods*, **17(11)**:1111-1117.

2. R. Schwessinger et al. (2020) DeepC: predicting 3D genome folding using megabase-scale transfer learning, *Nature methods*, **17(11)**:1118-1124.

3. J. Zhou (2022) Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale, *Nature Genetics*, **54**:725–734.