Снижение уровня шума ChIP-seq данных с помощью глубокого обучения

Фарутин В. В.

Научный руководитель: д.ф.-м.н. Б. А. Новиков

Научный консультант: А. А. Шпильман

Санкт-Петербургская школа физико-математических и компьютерных наук

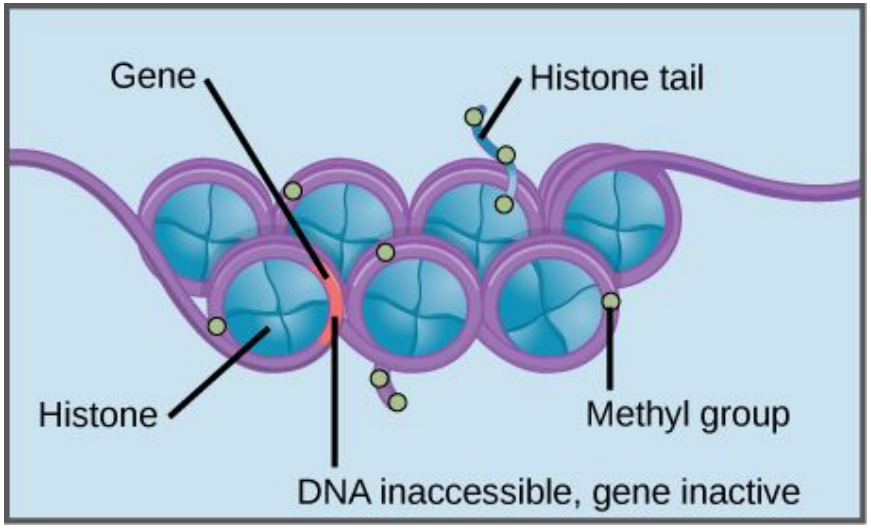
НИУ ВШЭ – Санкт-Петербург, 2020

Введение в область

Молекула ДНК накручена на нуклеосомы - группы белков гистонов (места связи - пики).

Гистоны участвуют в механизме эпигенетической регуляции.

Для реагирования на изменения окружающей среды организм меняет экспрессию генов, управляя гистонами.



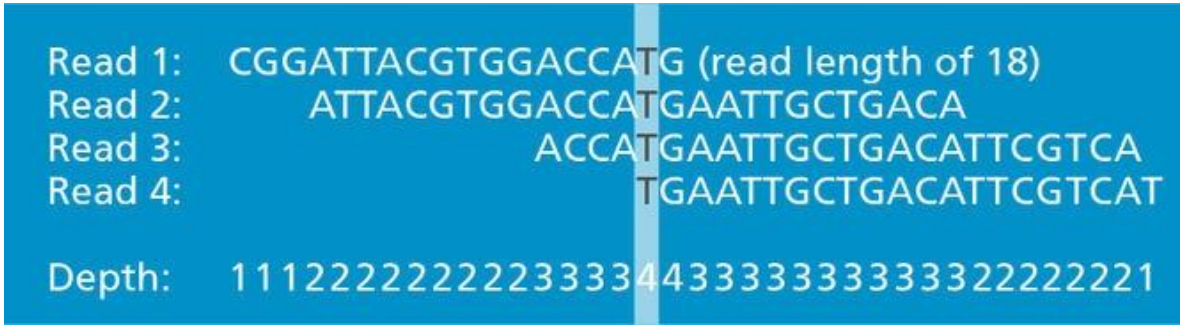
ChIP (chromatin immunoprecipitation) секвенирование:

- Дробление на фрагменты
- Выделение и осаждение нужных фрагментов с помощью антител
- Короткие секвенированные фрагменты упорядочиваются, для каждой позиции считается некоторая численная характеристика

Введение в область

Плохое качество данных:

1. Малая глубина секвенирования



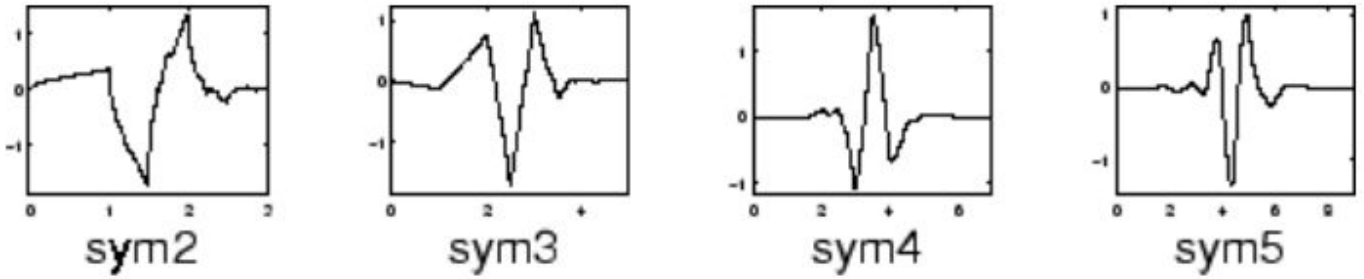
2. Малое количество клеток - Ultra Low Input ChIP-seq [1]:
   - Протокол способен работать при количестве клеток ~10^3-10^5 (обычный ChIP-seq использует ~10^7)
   - Имеет высокий уровень корреляции с качественным сигналом при делении на бины большого размера (2-4 kbp - base pairs)
   - Низкая точность на более детальном уровне

[1] J. Brind'Amour et al. "An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations". In: *Nat Commun 6, 6033* (2015)

Варианты решений: классический подход

Снижение уровня шума с помощью Wavelets [2][3]:

- Сигнал раскладывается в сумму базисных вейвлет-функций
- Выбирается порог для фильтрации слагаемых с маленькими коэффициентами



Решение уступает по точности методам машинного обучения.

[2] David L. Donoho. “Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data”. In: *Proceedings of Symposia in Applied Mathematics, 173-205* (1993).
[3] Ç. P. Dautov and M. S. Özerdem. "Wavelet transform and signal denoising using Wavelet method". In: *2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 1-4* (2018).

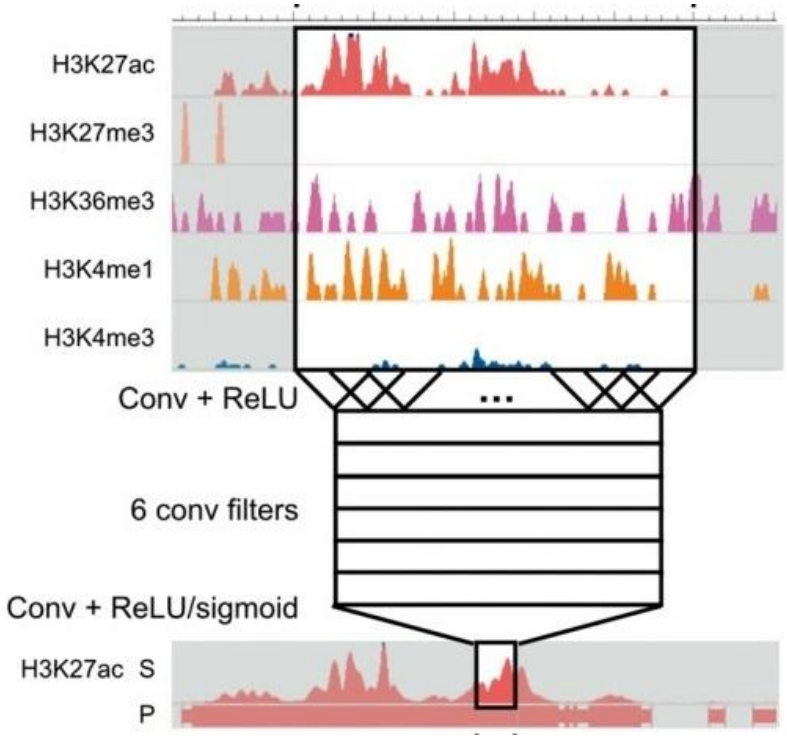
Варианты решений: глубокое обучение

Решение аналогичных проблем, но для других данных, простые модели сетей:

- ATAC-seq [4] - последовательность блоков с остаточными связями, формат данных совпадает с ChIP-seq
- scRNA-seq [5] - 3 слоя 64-32-64, данные представлены в виде матриц, а не последовательностей

ChIP-seq данные - **Coda** [6]:

- Небольшая сверточная сеть
- Одна архитектура для задачи предсказания сигнала и пиков на данных с разным шумом



[4] Avantika Lal et al. “AtacWorks: A deep convolutional neural network toolkit for epigenomics”. (2019)
[5] G. Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. In: *Nat Commun 10, 390* (2019).
[6] Koh Pang et al. “Denoising genome-wide histone ChIP-seq with convolutional neural networks”. In: *Bioinformatics. 2017;33(14):i225–i233* (2017).

Цель и задачи

Цель:

Снизить уровень шума в низкокачественных ChIP-seq данных, используя современные методы и подходы глубокого обучения.

Задачи:

- Разработать и реализовать модель для улучшения качества ChIP-seq данных
- Применить модель к данным с различными источниками шума
- Сравнить результаты с существующими решениями

Данные

Было использовано два датасета:

- Данные для 5 гистонов для клеток одинакового типа от людей с разными корнями для нескольких хромосом [7] - для экспериментов с малой глубиной секвенирования
- ULI-ChIP-seq датасет [1] с данными для 3 гистонов для клеток одинакового типа для нескольких хромосом - для экспериментов с малым количеством клеток

Для каждого гистона сигнал на 22 хромосомах длины ~10^7-10^8 bp.

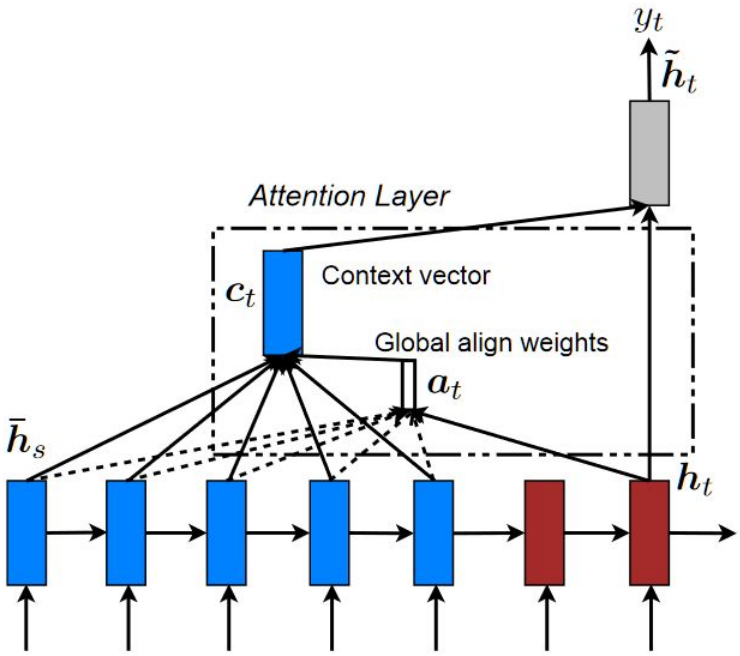
[1] J. Brind'Amour et al. "An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations". In: *Nat Commun 6, 6033* (2015)
[7] M. Kasowski et al. "Extensive Variation in Chromatin States Across Humans". In: *Science 2013 Nov 8; 342(6159):750-2* (2013).

Реализация модели: автоэнкодер

Автоэнкодер с использованием рекуррентных блоков:

- Кодировщик и декодировщик - LSTM
- Механизм внимания для решения проблемы ботлнека

Много вариантов конфигурации - различные виды внимания, функций подсчета весов, обучение с использованием teacher forcing и без.



[8] Ilya Sutskever et al. “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems. 2014. 09. Vol. 4.* (2014)

[9] Minh-Thang Luong et al. “Effective Approaches to Attention-based Neural Machine Translation”. (2015)

Результаты: автоэнкодер

Результаты для H3K27AC, обучение на клеточной линии GM12878, тестирование на GM18526 на хромосоме chr1 (9 970 000 bins):

	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.50	0.59	3.41	0.64
Coda	0.14	**0.85**	**0.44**	**0.83**
Моя модель (1)	**0.13**	0.85	0.60	0.82

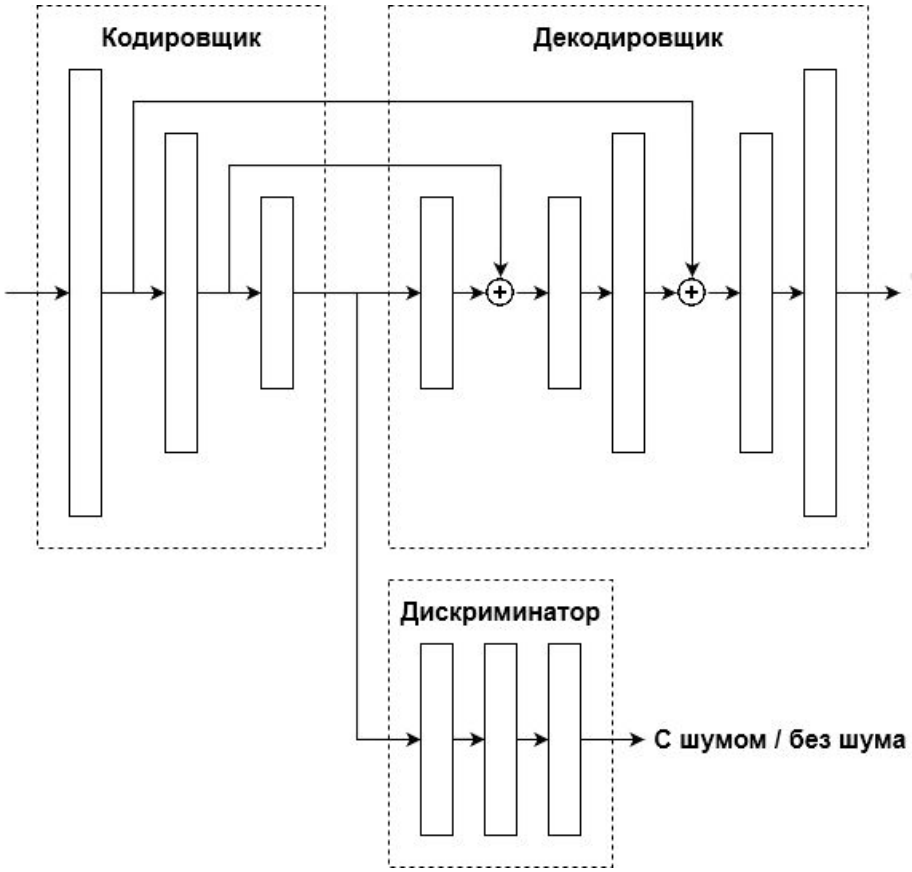
На длинных последовательностях сложно обучиться, на коротких недостаточно информации.

Реализация модели: сверточный автоэнкодер

U-net [10] подобная архитектура:

- Симметричная структура кодировщика и декодировщика
- Skip-connections между соответствующими слоями

Дискриминатор пытается отличить латентные вектора для чистого и грязного входа кодировщика.



[10] Olaf Ronneberger et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention / Springer. P. 234-241* (2015).
[11] Leslie Casas et al. “Adversarial Signal Denoising with Encoder-Decoder Networks”. (2019).

Результаты: сверточный автоэнкодер

Результаты для H3K27AC на данных с малой глубиной секвенирования:

	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.50	0.59	3.41	0.64
Wavelets	0.26	0.78	2.13	0.75
Coda	0.14	0.85	0.44	0.83
Моя модель (1)	0.13	0.85	0.60	0.82
Моя модель (2)	**0.09**	**0.90**	**0.43**	**0.85**

Результаты: данные с малой глубиной секвенирования

H3K4ME1	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.77	0.49	3.72	0.44
Coda	0.30	0.78	**0.47**	0.68
Моя модель (2)	**0.19**	**0.83**	0.56	**0.71**

H3K27ME3	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	1.17	0.21	2.22	0.17
Coda	**0.17**	0.68	0.20	0.36
Моя модель (2)	0.19	**0.72**	**0.18**	0.36

H3K4ME3	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.29	0.68	2.90	0.78
Coda	0.13	0.84	**0.50**	0.86
Моя модель (2)	**0.07**	**0.91**	0.62	**0.87**

H3K36ME3	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.86	0.45	3.80	0.32
Coda	0.12	0.89	0.24	0.72
Моя модель (2)	**0.09**	**0.92**	**0.19**	**0.73**

Результаты: Ultra Low Input данные

H3K4ME3, chr1	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.77	0.15	1.53	0.26
Coda	**0.47**	0.36	0.58	0.38
Моя модель (2)	0.53	**0.37**	**0.53**	**0.39**

H3K4ME3, chr2	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.62	0.17	1.47	0.34
Coda	**0.40**	0.37	0.60	0.43
Моя модель (2)	0.45	**0.39**	**0.54**	**0.45**

H3K4ME3, chr3	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.50	0.14	1.39	0.34
Coda	**0.37**	0.31	0.60	0.42
Моя модель (2)	0.42	**0.34**	**0.54**	**0.43**

H3K4ME3, chr4	MSE	Pearson	MSE, области пиков	Pearson, области пиков
Входные данные	0.61	0.18	1.50	0.35
Coda	**0.40**	0.40	0.62	0.44
Моя модель (2)	0.44	**0.42**	**0.55**	**0.45**

Выводы

- Реализовано несколько архитектур автоэнкодеров для снижения уровня шума в ChIP-seq данных
- Реализованные модели успешно справляются с задачей улучшения качества сигнала на данных с разными типами шума
- На основе экспериментов сверточный автоэнкодер с дискриминатором выбран в качестве финальной модели
- Финальная модель превосходит существующие решения (до -40% MSE и +6% Pearson по сравнению с ближайшим аналогом на всем сигнале)

Репозиторий: https://github.com/VadimFarutin/coda