

Предсказание активности и рациональный дизайн регуляторных районов генома и транскриптома эукариот методами машинного обучения

Пензар Д.Д.^{1,2*}, Ногина Д.Д.¹, Носкова Е.О.¹, Власов А.В.¹, Зинкевич А.О.¹, Кулаковский И.В.³

¹ Московский государственный университет им. М.В. Ломоносова, Москва, Россия

² Институт искусственного интеллекта AIRI, Москва, Россия

³ Институт белка РАН, Пущино, Россия

* dmitrypenzar1996@gmail.com

Ключевые слова: регуляторная геномика; глубокое машинное обучение; сверточные сети; регуляторные варианты; рациональный дизайн

Мотивация и цель: Исследования по расшифровке структуры и вычислительному предсказанию активности регуляторных районов генома эукариот сегодня развиваются особенно бурно благодаря синергии высокопроизводительных экспериментальных методов и машинного обучения. Накапливается все больше данных массовых параллельных экспериментов с репортерами, которые позволяют оценить активность сотен тысяч и даже миллионов регуляторных элементов. Неотъемлемым инструментом для работы с такими данными является машинное обучение, в особенности глубокие нейронные сети. Ранее мы предложили адаптацию архитектуры EfficientNetV2 [1], используемой для классификации изображений, к анализу нуклеотидных последовательностей. Предложенная нами модель LegNet [2] с большим отрывом победила конкурирующие решения на основе рекуррентных сетей и сетей на основе механизма внимания в задаче предсказания экспрессии репортерного белка в клетках дрожжей по последовательности промотора.

Сегодня нам удалось адаптировать модель к широкому спектру геномных задач: от предсказания активности регуляторных последовательностей в различных типах клеток человека до оценки эффекта однонуклеотидных замен в них и рационального дизайна регуляторных последовательностей с заданными свойствами.

Методы и алгоритмы: LegNet – полностью конволюционная нейронная сеть для работы с нуклеотидными последовательностями [2], реализованная на фреймворке pytorch [3]. В ходе обучения нейронной сети использовался метод one-cycle learning rate schedule [4]. Для обучения нейросети использовали данные из работ [5–7], для каждой из задач архитектура адаптировалась исходя из размера обучающей выборки, предсказываемой переменной и размера последовательностей. Для валидации использовали данные из работ [5–10].

Результаты: LegNet показывает SOTA (state-of-the-art) результаты на задаче предсказания экспрессии по регуляторной последовательности для промоторов дрожжей, регуляторных элементов дрозофилы, энхансеров человека. Для регуляторных районов генома человека достигнутое качество модели превосходит Enformer [11], предобученной на большом количестве экспериментов из базы

ENCODE. Модели, обученные для регуляторных районов генома человека, способны предсказать события аллель-специфичности по данным ADAstra [8] и UDACHA [9]. Также мы впервые предложили и успешно протестировали генеративную модель на основе диффузии для генерации промоторов с заданным уровнем экспрессии, не ограничиваясь работой с категориальными переменными [12, 13]. Наконец, мы адаптировали LegNet для работы на коротких последовательностях и, используя данные высокопроизводительного SELEX, показали применимость в задаче предсказания эффектов однонуклеотидных вариантов, измеренных в эксперименте SNP-SELEX [10].

Выводы: Нами была разработана новая архитектура сверточной нейронной сети LegNet и методика ее обучения, позволяющие достигнуть SOTA результатов на большом спектре задач регуляторной геномики. Предложенный метод применим для разных длин нуклеотидных последовательностей и разных организмов для решения задач предсказания активности регуляторных районов и для дизайна последовательностей с заданными свойствами.

Финансирование: Исследование было поддержано грантом РФФИ 20-74-10075.

Machine learning for rational design and reliable prediction of activity of gene regulatory regions

Penzar D.D.^{1,2*}, Nogina D.D.¹, Noskova E.O.¹, Vlasov A.V.¹, Zinkevich A.O.¹, Kulakovskiy I.V.³

¹ Lomonosov Moscow State University, Moscow, Russia

² AIRI, Moscow, Russia

³ Institute of Protein Research, RAS, Pushchino, Russia

* dmitrypenzar1996@gmail.com

Key words: regulatory genomics; deep learning; CNN rational design; regulatory variants

Motivation and Aim: Synergy of high-throughput experimental methods and machine learning has substantially driven the recent advances in deciphering the structure and predicting the activity of eukaryotic gene regulatory regions. The volume of available data obtained from massively parallel reporter assays is growing fast, and the already available dataset include activity estimates for hundreds of thousands or even tens of millions of diverse regulatory elements. Machine learning, in particular, deep neural networks became an essential tool to properly utilize such large-scale data for modeling gene regulatory regions. In previous studies, we had shown how to adapt the EfficientNetV2 architecture [1], originally used for image classification, to be effectively applicable for the analysis of nucleotide sequences. Our LegNet model [2] outperformed multiple competing solutions, including those based on recurrent networks and attention-based networks, in solving the problem of predicting reporter protein expression in yeast cells solely from the promoter sequence.

Here we show that LegNet is applicable to a wide range of genomic problems from predicting the activity of regulatory sequences in various human cell types to assessing the effect of single nucleotide variants and rational design of regulatory sequences with desired properties.

Methods and Algorithms: LegNet is a fully convolutional neural network optimized to handle nucleotide sequences [2] and implemented with PyTorch [3] using the one-cycle learning rate schedule for training [4]. Data from [5–7] was used for model training for different tasks, while the architecture was modified to account for varying sizes of the training set, the target variable, and the lengths of sequences. Data from [5–10] was used as independent test data.

Results: LegNet demonstrates state-of-the-art (SOTA) results in predicting expression from the regulatory sequence for yeast promoters, estimating the activity of *Drosophila* regulatory elements, and human enhancers. For human gene regulatory regions, the achieved model performance surpasses Enformer [11], which was pre-trained on a large number of experiments from the ENCODE database. Further, models trained for regulatory regions of the human genome are capable of predicting allele-specific events based on data from ADAstra [8] and UDACHA [9].

Also, we have proposed and tested a generative diffusion-based model for generating promoters with a desired level of expression, the first of its kind not limited by handling categorical variables only [12, 13]. Finally, we have adapted LegNet to shorter sequences and, with high-throughput SELEX data, demonstrated its capability in recognizing regulatory single nucleotide variants detected by the SNP-SELEX experiment [10].

Conclusion: We have developed a new convolutional neural network architecture, LegNet, and a training scheme allowing to achieve SOTA results in a wide range of machine learning applications in eukaryotic regulatory genomics. The proposed method is applicable to sequences of different lengths and for different organisms, enabling the prediction of regulatory region activity and the design of sequences with desired properties.

Funding: This study was supported by the Russian Science Foundation grant (20-74-10075 to I.V.K.)

Список литературы/References

1. Tan M., Le Q.V. EfficientNetV2: Smaller Models and Faster Training. [Preprint]. *Arxiv*. 2021. doi 10.48550/arXiv.2104.00298
2. Penzar D., Nogina D., Noskova E., Zinkevich A., Meshcheryakov G., Lando A., Rafi A.M., de Boer C., Kulakovskiy I.V. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics*. 2023;39(8):btad457. doi 10.1093/bioinformatics/btad457
3. Paszke A., Gross S. et al. Automatic differentiation in PyTorch. *Open Rev*. 2017. <https://openreview.net/pdf/25b8eee6c373d48b84e5e9c6e10e7cbbbce4ac73.pdf?ref=blog.prem.ai.io>
4. Smith L.N., Topin N. Super-convergence: very fast training of neural networks using large learning rates. [Preprint]. *Arxiv*. 2017. doi 10.48550/arXiv.1708.07120
5. Rafi A.M., Nogina D., Penzar D. et al. Evaluation and optimization of sequence-based gene regulatory deep learning models. *bioRxiv*. 2023. doi 10.1101/2023.04.26.538471
6. Agarwal V., Inoue F., Schubach M., Martin B.K., Dash P.M., Zhang Z., Sohota A., Noble W.S., Yardimci G.G., Kircher M., Shendure J., Ahituv N. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv*. 2023. doi 10.1101/2023.03.05.531189
7. de Almeida B.P., Reiter F., Pagani M. et al. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Gen*. 2022;54:613-624. doi 10.1038/s41588-022-01048-5
8. Abramov S., Boytsov A., Bykova D. et al. Landscape of allele-specific transcription factor binding in the human genome. *Nat Commun*. 2021;12:2751. doi 10.1038/s41467-021-23007-0
9. Buyan A. et al. Statistical framework for calling allelic imbalance in high-throughput sequencing data. *bioRxiv*. 2023. doi 10.1101/2023.11.07.565968
10. Yan J., Qiu Y. et al. Systematic analysis of binding of transcription factors to noncoding variants. *Nature*. 2021;591(7848):147-151. doi 10.1038/s41586-021-03211-0