doi 10.18699/bgrs2026-148

Изучение регуляторной грамматики НТО мРНК человека с помощью предсказательных и генеративных моделей глубокого обучения

Аристова Е.О.[1, 2]*, Хорошкин М.[3], Рыбаков А.К.[1], Пензар Д.Д.[2], Зинкевич А.О.[1], Кулаковский И.В.[2, 4]
[1] *Факультет биоинженерии и биоинформатики МГУ, Москва, Россия*
[2] *Институт общей генетики им. Вавилова РАН, Москва, Россия*
[3] *Университет Калифорнии, Сан-Франциско, США*
[4] *Институт белка РАН, Пущино, Россия*
* *aristova@vigg.ru*

Ключевые слова: регуляция экспрессии генов; тканевая специфичность; нетранслируемые области (НТО); диффузионные модели; РНК-мотивы; РНК-связывающие белки; посттранскрипционная регуляция; РНК-терапия

Мотивация и цель: Регуляция экспрессии генов у высших эукариот необходима для формирования и поддержания идентичности разнообразия клеточных типов, для обеспечения клеточного ответа на сигналы и изменение условий внешней среды, как на уровне отдельных клеток так и целого организма.
На транскрипционном уровне ткань-специфичнойэкспрессии генов активно изучаются и достаточно хорошо охарактеризованы. Однако, нельзя отрицать принципиальную роль и посттранскрипционных регуляторных процессов на уровне РНК — регуляции процессинга, трансляции и стабильности транскриптов. Ранее, используя данные массовых параллельных репортёрных экспериментов (МПРЭ), нами была разработана предсказательная модель «от последовательности к активности» для оценки клеточно-специфичной активности нетранслируемых областей (НТО) мРНК человека [1]. В настоящей работе мы используем эту модель глубокого обучения и лежащие в её основе данные МПРЭ чтобы оценить вклад отдельных регуляторных мотивов и исследовать возможности различных генеративных подходов *in silico* для создания новых 5’- и 3’-НТО с клеточно-специфичной активностью — то есть неоднородной эффективностью синтеза белка из мРНК под контролем НТО в разных типах клеток.
Методы и алгоритмы: В данном исследовании использовались данные крупномасштабного параллельного репортёрного анализа (МПРЭ), включающего 60 000 НТО (5’ и 3’), активность которых была измерена в 6 клеточных линиях человека [1].
Для характеристики известных и выявления новых регуляторных последовательных мотивов, определяющих активность НТО, мы применили три взаимодополняющих подхода. Во-первых, был проведён виртуальный эксперимент in silico МПРЭ: множественные копии известных РНК-мотивов встраивались в случайные НТО-остовы, а полученные последовательности подавались на вход предсказательной модели. Это позволило выявить активирующее или репрессирующее действие множества известных мотивов. Во-вторых, был применён метод интерпретации модели TF-MoDISco для извлечения паттернов, самостоятельно «выученных» моделью. В-третьих, использовался классический инструмент поиска РНК-мотивов FIRE для обнаружения паттернов, объясняющих дифференциальную активность НТО по данным исходного МПРЭ
Наконец, мотивы из всех трёх источников кластеризовались совместно с известными мотивами РНК-связывающих белков (РСБ) для установления связи между обнаруженными паттернами и выявления новых регуляторных элементов.
Для исследования возможностей различных генеративных подходов машинного обучения при конструировании НТО с клеточно-специфичной активностью были применены следующие методы: модели «холодной диффузии» и метод согласования потоков [2, 3], ориентированные на конкретный клеточный тип; генетический алгоритм с предсказателем, оптимизирующий разницу активностей между целевым и нецелевым клеточными типами; градиентная оптимизация (Fast SeqProp) [4], также нацеленная на максимизацию этой разницы; и состязательная генеративная модель (в стиле GAN, аналогичная UTRGAN) [5].
Результаты: С помощью предсказательной модели совместно с набором данных МПРЭ был количественно оценён вклад отдельных мотивов в активность НТО. Анализ in silico МПРЭ и прямой поиск мотивов показали, что модель воспроизводит множество известных регуляторных эффектов: репрессирующую роль коротких апстрим-рамок считывания в 5’-НТО, а также разнообразный репертуар мотивов, как дестабилизирующих транскрипты (Regnase-подобные, AU-богатые и Pumilio-подобные мотивы), так и многочисленные активирующие мотивы, распознаваемые стабилизирующими или усиливающими трансляцию РСБ. Интеграция результатов FIRE и TF-MoDISco позволила восстановить более четырёхсот известных мотивов РНК-связывающих белков и идентифицировать свыше трёхсот новых мотивов, в том числе несколько десятков комплексных паттернов, не имеющих аналогов в существующих базах данных мотивов РСБ или миРНК.
Далее была проведена сравнительная оценка генеративных методов, выявившая различные компромиссы между функциональными и последовательностными характеристиками. Диффузия и метод соглаосвания потока порождали последовательности, по составу близкие к геномным, при этом обладающие высокой новизной (то есть, характерно непохожие на исходные фрагменты нативных геномных НТО из обучающей выборки) и достаточно сложные (с большим числом уникальных подстрок). Генетический алгоритм

сохранял новизну и сложность последовательностей и успешно создавал тканеспецифичные дизайны, однако приводил к смещению GC-состава. GAN-сгенерированные последовательности, напротив, оказались чрезмерно похожи на последовательности из обучающей выборки, что свидетельствует об ограниченном исследовании пространства за пределами нативных последовательностей. Наконец, градиентная оптимизация нередко порождала последовательности низкой лсожности и композиционно смещённые по составу [2–5]. В совокупности эти результаты послужили основанием для приоритизации дизайнов на основе диффузии и генетического алгоритма как взаимодополняющих генераторов НТО, совместно обеспечивающих баланс между специфичностью, реалистичностью и разнообразием последовательностей. Лучшие дизайны были успешно валидированы в независимых экспериментах в люциферазном репортерном эксперименте.

Выводы: Анализ мотивов вскрыл ключевые компоненты посттранскрипционной регуляторной грамматики на уровне последовательности: сотни известных мотивов были точно воспроизведены, а многие новые регуляторные элементы продемонстрировали выраженное влияние на активность НТО.

На уровне дизайна последовательностей наиболее эффективными методами оказались генеративные модели на основе диффузии и генетические алгоритмы с предсказателем: совместно они обеспечивают значительную разницу активностей между целевым и нецелевым клеточными типами при сохранении реалистичного нуклеотидного состава, высокой новизны и сложности последовательностей, что делает их особенно привлекательными инструментами для рационального конструирования НТО. В целом, растущая способность подходов машинного обучения к конструированию и точной оптимизации последовательностей НТО открывает широкие перспективы для совершенствования мРНК-терапии, включая мРНК-вакцины, обеспечивая тонкий контроль над эффективностью трансляции и минимизацию нецелевой экспрессии, способной влиять на безопасность и точность иммунного ответа.

Финансирование: Исследование частично поддержано за счет государственного гранта (№ 125091010189-3).

Exploring the regulatory grammar of human UTRs with predictive and generative deep learning models

Aristova E.O.[1, 2*], Khoroshkin M.[3], Rybakov A.K.[1], Penzar D.D.[2], Zinkevich A.O.[1], Kulakovskiy I.V.[2, 4]

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[2] *Vavilov Institute of General Genetics, RAS, Moscow, Russia*

[3] *University of California, San Francisco, CA, USA*

[4] *Institute of Protein Research, RAS, Pushchino, Russia*

** aristova@vigg.ru*

Key words: gene expression regulation; tissue specificity; untranslated regions; UTRs; diffusion models; RNA motifs; RNA-binding proteins; post-transcriptional regulation; RNA therapeutics

Motivation and Aim: In higher eukaryotes, gene expression regulation is essential for establishing and maintaining the diversity of cell types as well as responses to signals and environmental conditions at the level of individual cells and the whole organism.

While transcriptional mechanisms of tissue- and cell type-specificity are actively studied and well understood, recent findings emphasize the essential role of post-transcriptional RNA-level regulatory processes, including RNA processing, translation, and stability. Recently, using rich data from massively parallel reporter assays, we developed a sequence-to-function predictive model to accurately evaluate the cell type-specific activity of untranslated regions (UTRs) of human mRNAs [1]. Here, we reuse this deep learning model and the underlying MPRA data to assess the contributions of individual regulatory motifs and explore the capabilities of different generative in silico approaches in creating novel 5’ and 3’ UTRs to achieve cell type-specific activity, that is, the non-uniform efficiency protein production from UTR-controlled mRNAs in different cell types.

Methods and Algorithms: In this study, we utilized the data from a massive parallel reporter assay (MPRA) comprising 60,000 UTRs (5’ and 3’) whose activity was profiled in 6 distinct human cell lines [1].

To characterize known and identify novel regulatory sequence motifs underlying UTR activity, we combined three complementary strategies. First, we performed a virtual in silico MPRA experiment by embedding multiple copies of known RNA motifs into random UTR backbones and supplying the resulting sequences to the predictive deep learning model. By doing so, we revealed the known motifs’ impacts as activators or repressors. Second, we applied a model interpretation technique TF-MoDISco to extract the sequence patterns learned by the predictive model. Third, we applied the classic RNA motif discovery tool, FIRE, to detect sequence patterns explaining differential UTR activity observed in the original MPRA. Finally, motifs from all three sources were clustered together with known motifs of RNA-binding proteins to link the discovered patterns to reveal novel regulatory elements.

Furthermore, to explore the capabilities of different generative machine learning techniques in design of the UTRs with cell type-specific activity, we employed the following generative approaches: “cold diffusion” [1, 2] and flow-matching [3] models targeting a particular cell type; a predictor-assisted genetic algorithm optimizing activity differences between the target and off-target cell types; gradient-based optimization in the spirit of Fast SeqProp [4] also optimizing differences between the target and off-target cell types; and an adversarial generative model (GAN-style, analogous to UTRGAN [5]).

Results: Using a predictive model together with the MPRA dataset, we quantified how individual motifs affect UTR activity. In silico MPRA and motif discovery analyses showed that the model captured multiple known regulatory effects, including the uORFs repressing role in 5' UTRs, and a diverse repertoire of decay-promoting motifs (Regnase-like, AU-rich, and Pumilio-like sites) and many activating motifs recognized by stabilizing or translation-enhancing RBPs. Integrating FIRE and TF-MoDISco results recovered more than four hundred known motifs of RNA-binding proteins and identified more than three hundred high-confidence novel motifs, including four dozen high-complexity patterns not matching anything available in current RBP or miRNA databases.

Next, we explored the generative methods, revealing distinct trade-offs between functional and sequence metrics. Diffusion and flow matching produced compositionally similar to genomic sequences that were highly novel (i.e., non-similar to the initial fragments of native genomic training UTRs assessed in the MPRA) and sufficiently complex. The genetic algorithm preserved novelty and complexity and yielded tissue-specific designs but showed moderate GC content drift. In contrast, GAN-generated sequences were overly similar to those present in the training set, indicating limited exploration beyond the native sequence space. Finally, gradient-based optimization often yielded low-complexity, compositionally biased outputs [2–5]. Together, these results motivated prioritizing diffusion- and genetic algorithm-based designs as complementary UTR generators that jointly balance specificity, realism, and sequence diversity. In the end, these designs successed in the independent experimental validation in luciferase assays.

Conclusion: Motif analysis revealed important components of the sequence-level post-transcriptional regulatory grammar, with hundreds of known motifs accurately recovered and many novel regulatory sequence elements with strong effects on UTR activity.

At the level of sequence design, diffusion-based generators and predictor-guided genetic algorithms emerged as the most effective methods, jointly achieving large target–off-target activity gaps while maintaining realistic nucleotide composition, high sequence novelty and complexity, making them particularly attractive tools for rational UTR engineering in mRNA-based applications. All in all, the growing ability of machine learning approaches to accurately optimize UTR sequences holds significant promise for improving mRNA-based therapeutics, including mRNA vaccines, by enabling fine-grained control over translation efficiency and minimizing off-target expression effects that may influence safety and immune response precision.

Funding: This study was partially supported by state assignment 125091010189-3.

Список литературы/References

1. Khoroshkin M., Zinkevich A., Aristova E., Yousefi H., Lee S.B., Mittmann T., Manegold K., Penzar D., Raleigh D.R., Kulakovskiy I.V., Goodarzi H. A generative framework for enhanced cell-type specificity in rationally designed mRNAs. bioRxiv. 2024 Dec 31. doi:10.1101/2024.12.31.630783.
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. Lipman Y., Chen R.T.Q., Ben-Hamu H., Nickel M., Le M. Flow Matching for Generative Modeling. arXiv:2210.02747. 2022.
4. Linder J., Seelig G. Fast activation maximization for molecular sequence design. BMC Bioinformatics. 2021;22:510. doi:10.1186/s12859-021-04437-5.
5. Barazandeh S., Ozden F., Hincer A., Seker U.O.S., Cicek A.E. UTRGAN: learning to generate 5′ UTR sequences for optimized translation efficiency and gene expression. Bioinform Adv. 2025;5(1):vbaf134. doi:10.1093/bioadv/vbaf134.