

# Предсказание класса длинных некодирующих РНК по расположению в геноме методами глубокого машинного обучения

В.С. Соколов<sup>1</sup>, А.Ю. Пронозин<sup>2</sup>, Д.А. Афонников<sup>1,2</sup>

<sup>1</sup>Новосибирский национальный исследовательский государственный университет

<sup>2</sup>Институт цитологии и генетики Сибирского отделения Российской академии наук

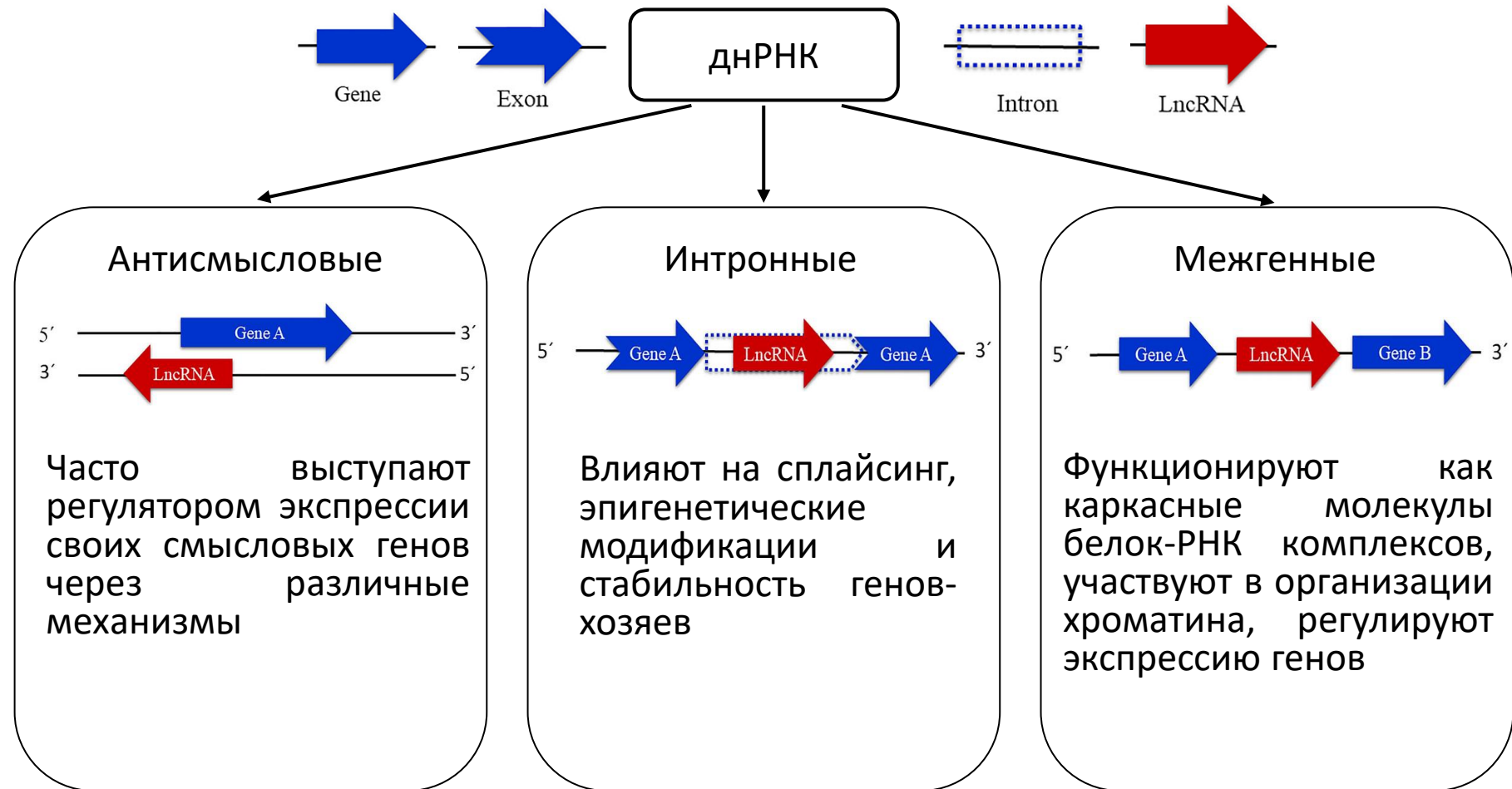
Новосибирск, 2025

# Введение

Длинные некодирующие РНК (днРНК) – РНК, не кодирующие белки, длиной более 200 нк

## Свойства днРНК:

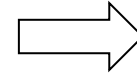
- Составляют более 50% всех РНК экспрессируемых в клетке
- Для некоторых днРНК показаны их регуляторные функции в геноме
- От класса днРНК зависят многие её функции
- Слабо консервативны
- Слабо изучены



# Данные

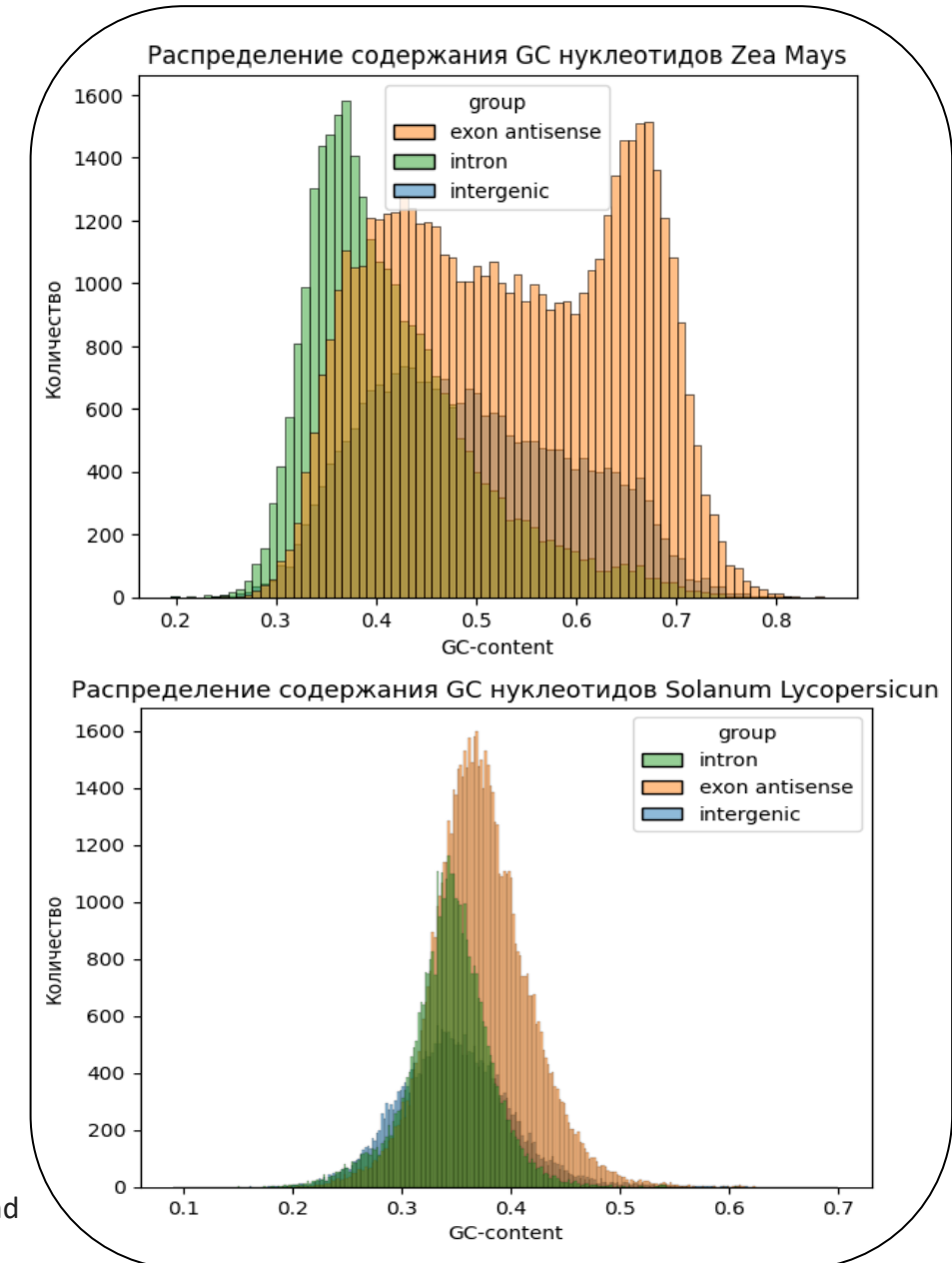
Набор данных получен выравниванием на геном транскриптов и определении их местоположения относительно генов [\*].

Последовательности видов различаются по GC-составу и разделяются на две группы по типу распределения.



В наборе данных представлено 5 видов с около 100000 последовательностями для каждого вида.

| Вид                  | Кол-во днРНК | Средняя длина днРНК | Средний GC-состав | Доля днРНК длиннее 512 |
|----------------------|--------------|---------------------|-------------------|------------------------|
| Hordeum              | 143279       | 469                 | 0.50              | 0.29                   |
| Oriza Sativa         | 70901        | 393                 | 0.46              | 0.18                   |
| Solanum Lycopersicun | 139410       | 568                 | 0.36              | 0.38                   |
| Solanum Tuberosum    | 61935        | 501                 | 0.37              | 0.25                   |
| Zea Mays             | 105811       | 391                 | 0.49              | 0.19                   |



\* Pronozin, A.Y.; Afonnikov, D.A. IAnnoLncRNA: A Snakemake Pipeline for a Long Non-Coding-RNA Search and Annotation in Transcriptomic Sequences. *Genes* 2023, 14, 1331. <https://doi.org/10.3390/genes14071331>

# Методы

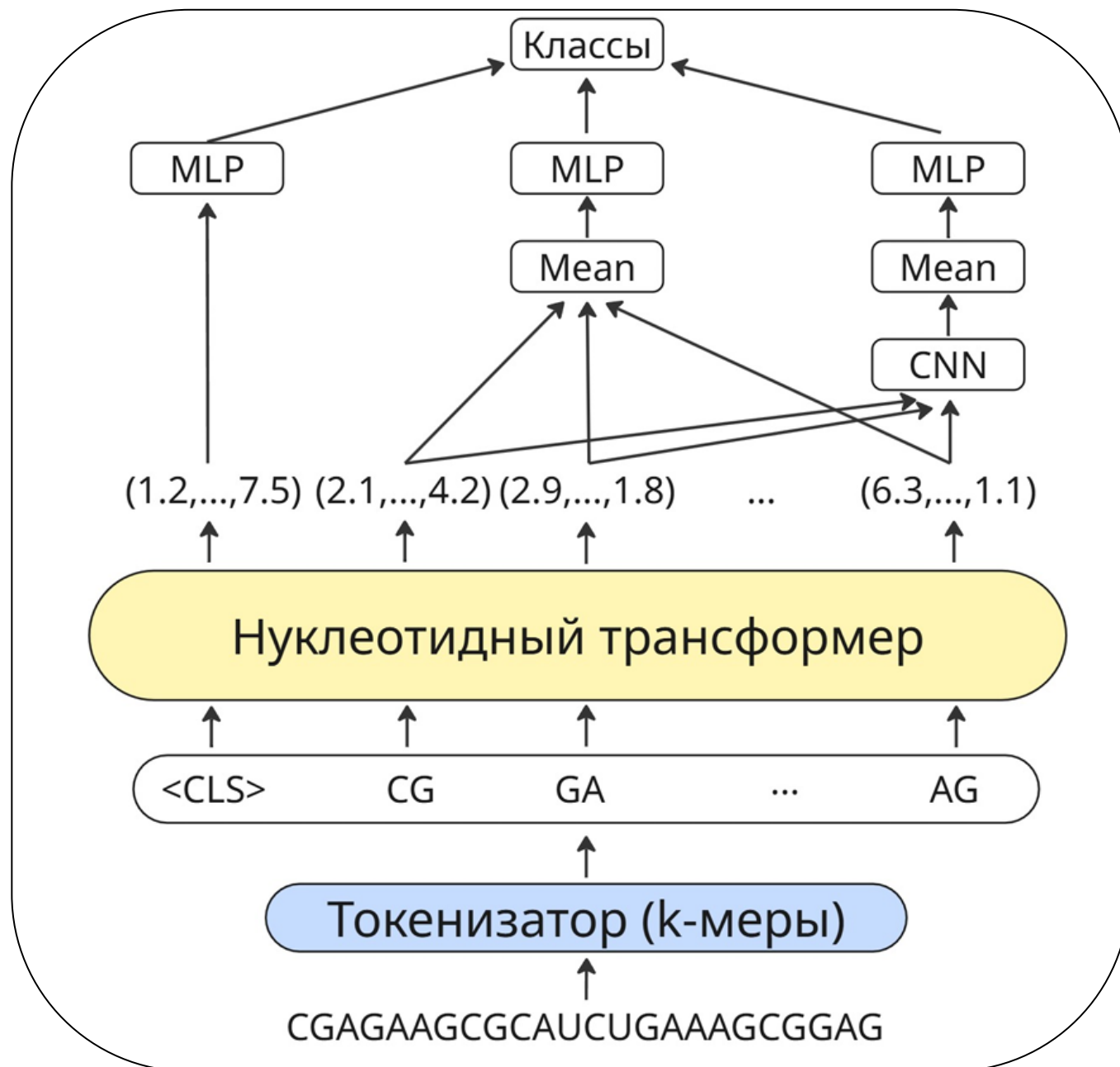
В работе рассмотрены различные предобученные нуклеотидные трансформерные модели на основе нейронных сетей.

Для каждой модели рассмотрены 3 способа агрегации информации:

- полносвязная нейронная сеть на выходе <CLS> токена,
- полносвязная нейронная сеть с усреднением по всем выходам,
- свёрточная нейронная сеть с усреднением по всем выходам.

Реализован подход PLEK [\*], использующий частоты k-мер и методы машинного обучения для классификации РНК.

\* Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014 Sep 19;15(1):311. doi: 10.1186/1471-2105-15-311. PMID: 25239089; PMCID: PMC4177586



# Результаты

🧬 Лучший нуклеотидный трансформер ERNIEBERT с длинным контекстным окном – 2048.

🧬 Лучший подход агрегации информации с выходов трансформерной модели – подход на основе свёрточных нейронных сетей. Свёрточная нейронная сеть лучше распознаёт локальные признаки – мотивы в последовательности.

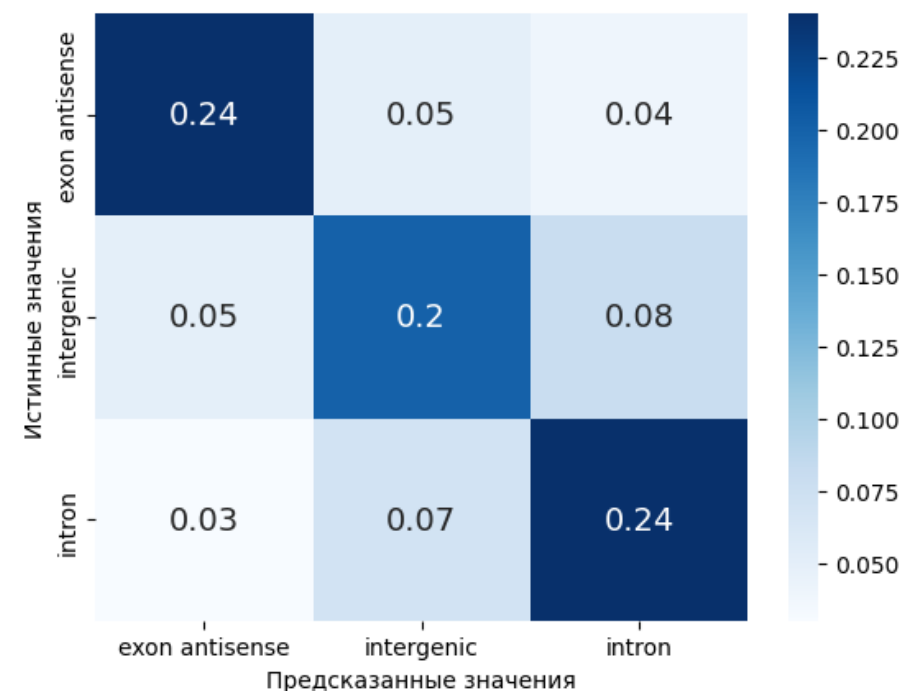
Точность классификации разными методами по F1 метрике для каждого класса

| Базовая модель | <CLS><br>ERNIE<br>1028 | Mean<br>ERNIE<br>1028 | CNN<br>ERNIE<br>1028 | Catboost    |
|----------------|------------------------|-----------------------|----------------------|-------------|
| Антисмысловые  | 0.70                   | <b>0.72</b>           | <b>0.72</b>          | <b>0.72</b> |
| Межгенные      | 0.56                   | 0.56                  | <b>0.61</b>          | 0.60        |
| Интронные      | 0.63                   | 0.65                  | <b>0.67</b>          | <b>0.67</b> |
| F1-macro       | 0.63                   | 0.65                  | <b>0.67</b>          | <b>0.67</b> |

🧬 Метод PLEK с использованием градиентного бустинга Catboost показал результаты, сравнимые с нейросетевыми подходами.

🧬 Наилучшее качество предсказания имеет класс антисмысловых днРНК, что связано с их большей консервативностью по сравнению с другими типами.

Матрица ошибок



# Выводы

- днРНК разных видов и классов различаются по длине и GC-составу
- Подход со свёрточной нейронной сетью для предобученного трансформера показал лучшие результаты
- Точность классификации варьируется в зависимости от класса днРНК
- С наибольшей точностью предсказываются антисмысловые днРНК, с наименьшей - интронные
- Наибольшую точность показали методы CNN ERNIE BERT с контекстом длиной 2048 и градиентный бустинг с подготовкой признаков методом PLEK

# Планы

- Добавление дополнительных признаков, извлечённых из последовательности (например, вторичная структура)
- Добавление эволюционной информации из множественного выравнивания
- Исследование точности предсказания моделей для разных видов
- Объяснение результатов
- Интерпретация выученных правил моделями для классификации