Факультет
Компьютерных наук

Департамент Программной
Инженерии

Москва
2024

Идентификация вторичных структур ДНК с помощью нейросетей

AI-powered Identification of DNA Secondary Structures

Тип проекта: исследовательский
Вид проекта: индивидуальный

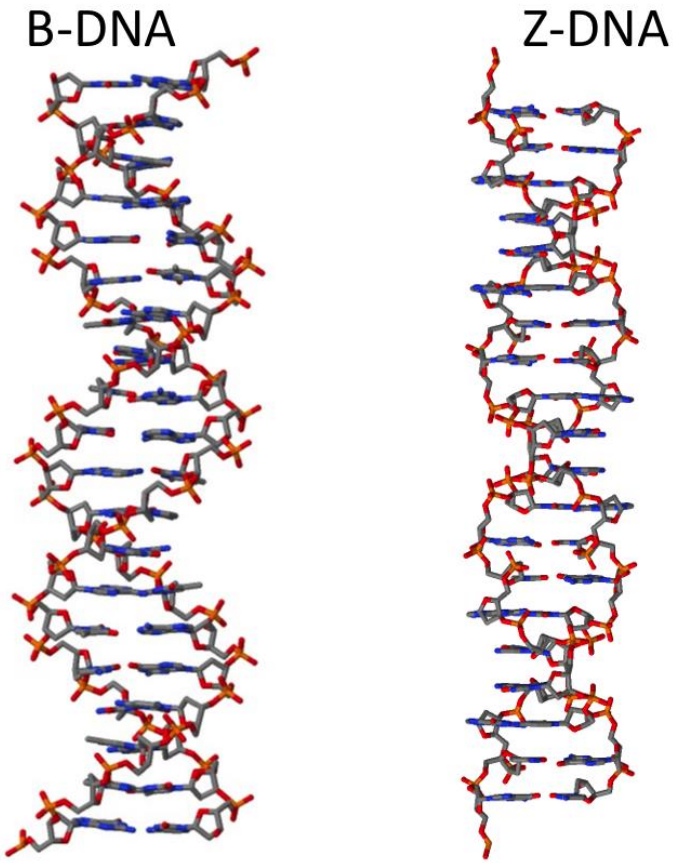
Автор проекта: Лаптева Анна, БПИ215
Руководитель проекта: Преподаватель НИУ ВШЭ, бакалавр, **Боревский Андрей Олегович**

Предметная область. Z-ДНК и B-ДНК

Помимо всем известной стандартной правой формы ДНК (B-ДНК), в 1979 году была обнаружена левосторонняя форма, называемая Z-ДНК.

Функциональная роль Z-ДНК разнообразна, и многие из её аспектов остаются недостаточно исследованными.

Установлено, что Z-ДНК может действовать как репрессор в промоторе гена ADAM-12, который известен своей чрезмерной активностью во многих формах рака у человека. Она также обнаружена в области гиппокампа мозговых образцов, сильно пораженных болезнью Альцгеймера. Белки ADAM, содержащие Z-ДНК в промоторной области, ассоциированы с различными метаболическими и воспалительными заболеваниями, включая диабет, сепсис, болезнь Альцгеймера и ревматоидный артрит.



Цель и задачи работы

Цель работы – разработка модели глубинного обучения, способной с высокой точностью предсказывать участок Z-ДНК на основе вторичных структур ДНК.

В рамках данной работы должны быть выполнены следующие задачи:

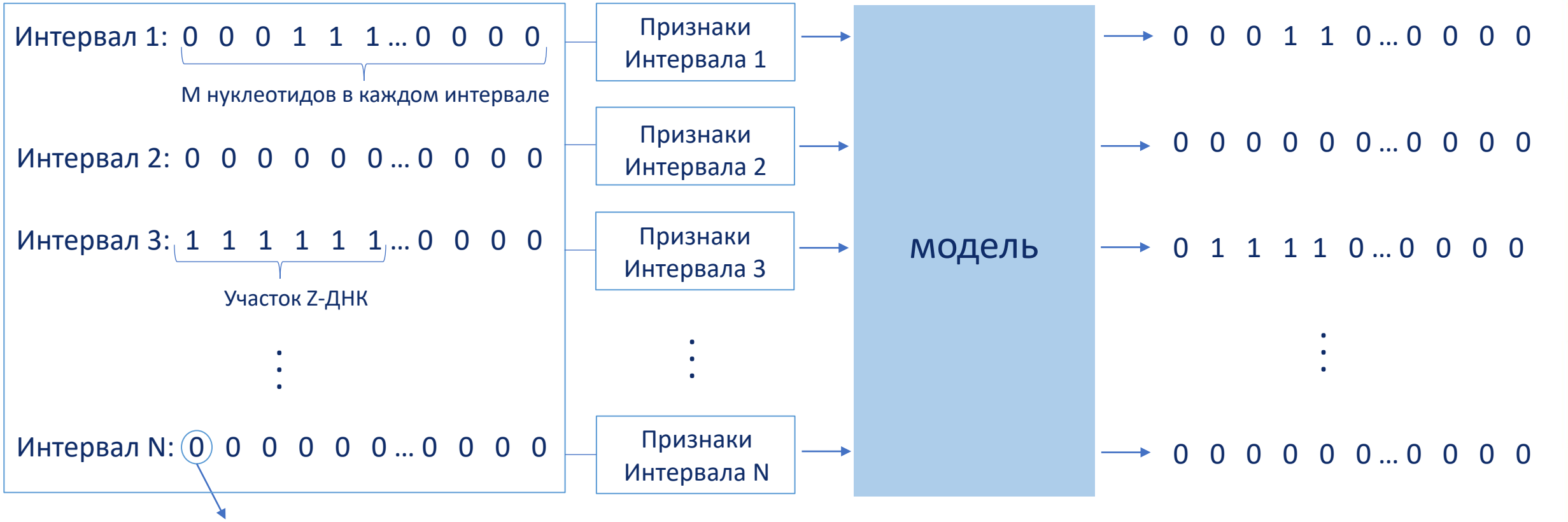
1) Реализация и оптимизация модели глубинного обучения, разработанной научно-учебной лабораторией искусственного интеллекта для вычислительной биологии НИУ ВШЭ [1].
2) Исследование эффективности различных архитектур моделей глубинного обучения для задач исследования ДНК и вторичных структур ДНК.
3) Реализация модели глубинного обучения, параметры которой можно проинтерпретировать в будущем.

Актуальность работы

Актуальность работы заключается в разработке эффективной модели глубинного обучения для предсказания участков Z-ДНК для получения инструмента для диагностирования участков Z-ДНК.

Аномальное образование участков Z-ДНК ассоциируется с рядом серьезных заболеваний. Среди них могут быть рассмотрены болезни, связанные с нарушениями структуры и функционирования ДНК, такие как рак, нейродегенеративные заболевания, аутоиммунные расстройства и другие патологии, что подчеркивает важность точного и надежного определения участков Z-ДНК.

Формат данных



Каждый нуклеотид имеет 1058 признака, анализируя признаки каждого нуклеоида в интервале, модель предсказывает, относится ли нуклеотид к Z-ДНК (выдает 1) или нет (выдает 0)

Анализ существующих решений

Лучшая модель лаборатории DeepZ, полученная после проведения более 150 экспериментов с различными архитектурами, основана на LSTM.

Метрики DeepZ на интервале размером 5000 нуклеотидов:

ROC-AUC = 85%, F1-score = 39.1%

Модель имеет достаточно низкий F1-score, поэтому первостепенной задачей стоит увеличение F1-score, и в дальнейшем для сравнение моделей будем опираться больше на значение этой метрики.

```
class DeepZ(nn.Module):
    def __init__(self):
        super().__init__()
        self.rnn = nn.LSTM(1058, 500, 2, bidirectional=True)
        self.seq = nn.Sequential(
                    nn.Dropout(0.5),
                    nn.Linear(2 * 500, 500),
                    nn.Sigmoid(),
                    nn.Dropout(0.5),
                    nn.Linear(500, 2)
        )

    def forward(self, x):
        x, (h_n, c_n) = self.rnn(x)
        x = self.seq(x)
        return F.log_softmax(x, dim=-1)
```

Метрики качества

В качестве основных метрик качества были выбраны метрики ROC-AUC и F1-score, так метрика F1-score устойчив к дисбалансу классов, а ROC-AUC оценивает качество классификатора в целом

$$recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$