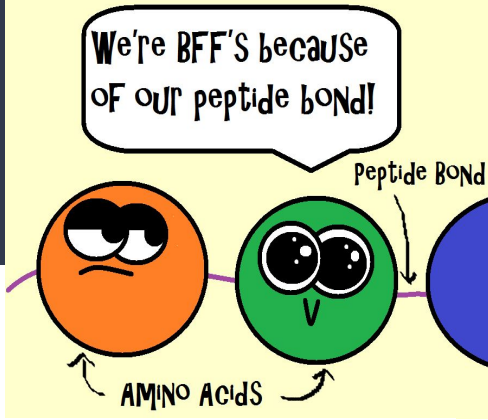
Предсказание специфичности А доменов в синтетазах нерибосомных пептидов

Бурцев Артём
Камелин Алексей
Кольцов Семён

Научный руководитель:
Азат Тагирджанов

NRP – нерибосомные пептиды



Синтезируются не рибосомами, а специальными белками (у бактерий и грибов)

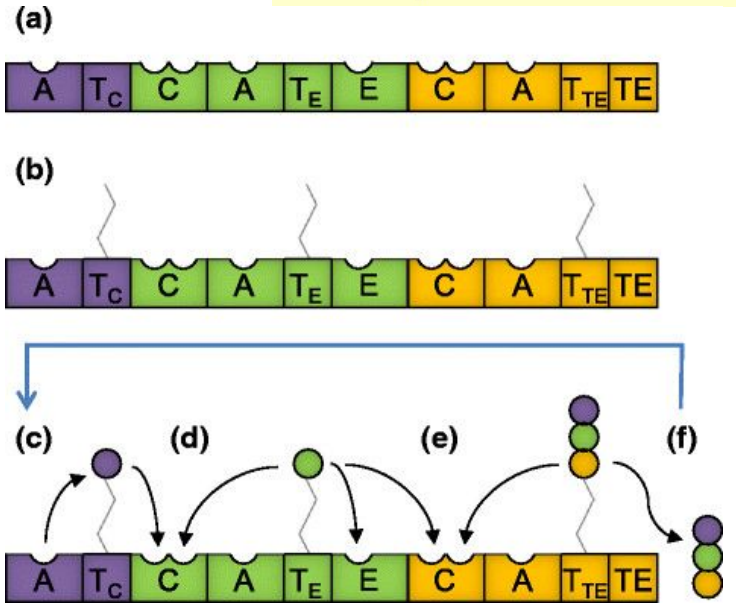
Протеиногенные аминокислоты:

20 хорошо нам знакомых α-аминокислот

В состав NRP входят:

Протеиногенные аминокислоты +

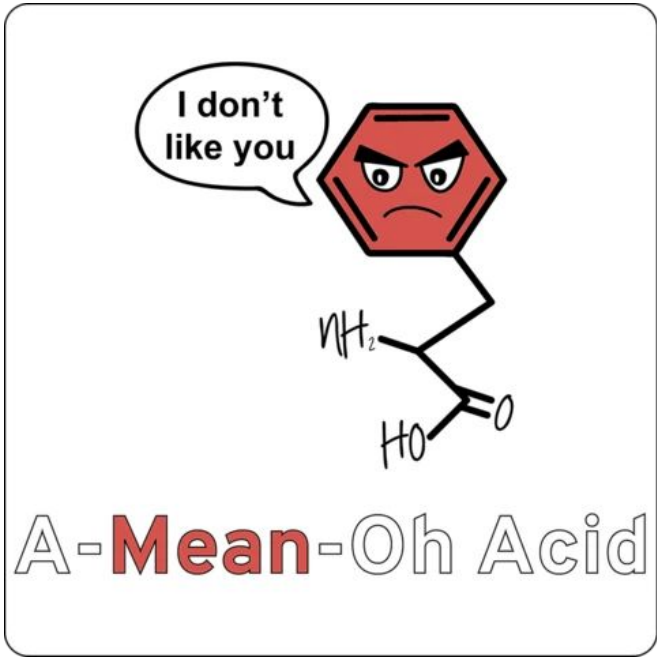
{aad, abu, b-ala, b-lys, bht, dab, очень много}



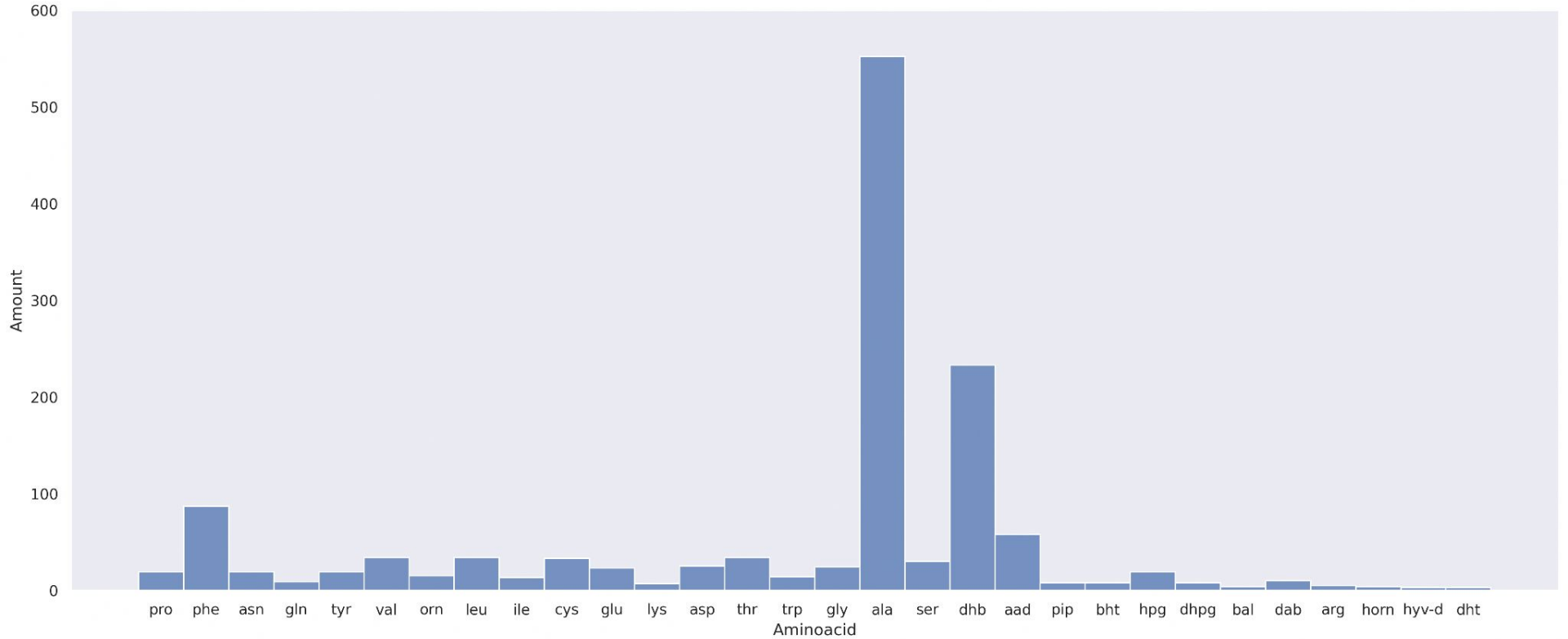
Цель: NRP: A-domain …ACGGTCA… → ML → Glycine

Задачи:

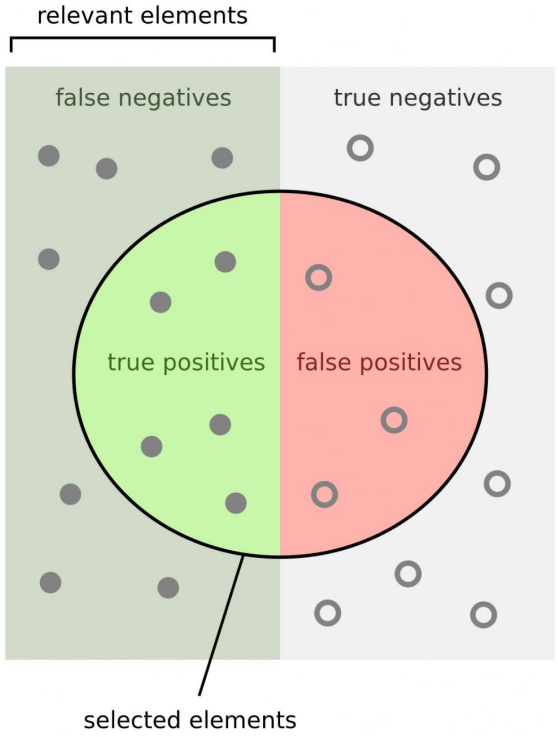
- Синхронизация баз данных
- Предсказания классическими ML
- Предсказание скрытыми марковскими моделями
- Deep learning



Распределение аминокислот в данных



Метрики в машинном обучении



Accuracy: Right answers / (Right + wrong answers)

How many selected items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

How many relevant items are selected?

Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

F1-score:
2*((precision*recall)/(precision+recall))

Наивный байес и логистическая регрессия

- **Наивный Байес:**
  - Accuracy на кросс-валидации: 0.875 ± 0.035
  - f1-score-макро-метрика: 0.691

- **Логистическая регрессия:**
  - Accuracy на кросс-валидации: 0.858 ± 0.034
  - f1-score-макро-метрика: 0.691

Неслучайный лес и метод опорных векторов

- **Случайный лес:**
  - Accuracy на кросс-валидации: 0.836 ± 0.019
  - f1-score-макро-метрика: 0.708

- **Метод опорных векторов:**
  - Accuracy на кросс-валидации: 0.870 ± 0.036
  - f1-score-макро-метрика: 0.708

Сравнение с NRPSpredictor2

NRPSpredictor2 - враги

Результаты тестирования на нашем датасете:

NRPSpredictor2

- Accuracy-метрика: 0.79
- f1-score-макро-метрика: 0.57

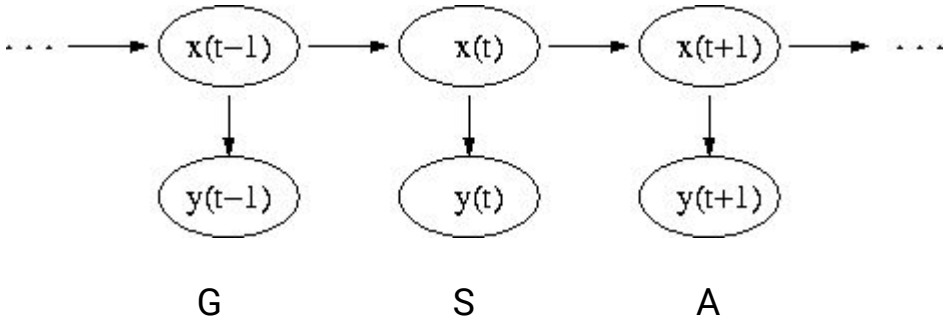
Скрытые марковские модели

```
Accuracy:

0.85 +- 0.01

F1-score:

0.72 +- 0.03
```



Искусственный интеллект

Нейронные сети:

1) Dense: 0.8 +- 0.03
2) Convolutional (свёрточная): 0.77 +- 0.03
3) Residual CNN: 0.76 +- 0.04

Результаты

Лучшая точность:

Наивный байесовский классификатор (0.875)

Мы молодцы

The End