

# Предсказание вторичной структуры белка на основе методов Deep Learning

Караваева Валерия

Институт биоинформатики

Руководитель — И. Дрокин



Санкт-Петербург  
2015г.

## Мотивация изучения белков

- Изучение функций белка.
- Моделирование взаимодействия с другими белками.
- Изучение взаимодействия белков с лекарственными препаратами и др.



## Задача

*Предсказание вторичной структуры белка с использованием первичной структуры.*

## Задача

*Предсказание вторичной структуры белка с использованием первичной структуры.*

- *Способ решения: использование методов глубоких рекуррентных сетей (RNN).*

## Задача

*Предсказание вторичной структуры белка с использованием первичной структуры.*

- *Способ решения: использование методов глубоких рекуррентных сетей (RNN).*
- *Причина: RNN «копирует» реальный процесс синтеза белка.*

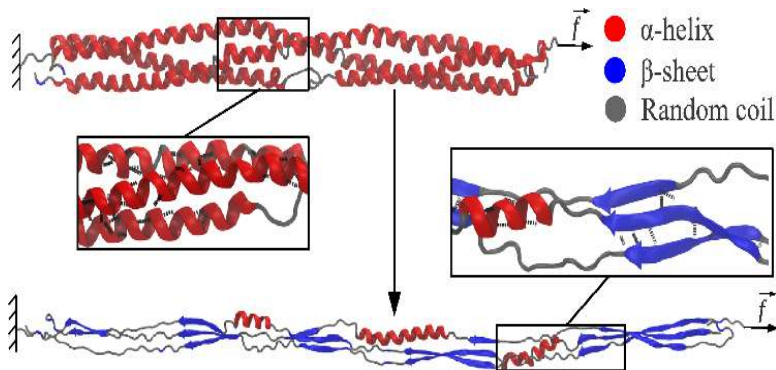
## Формализация задачи:

- $P = a_1, \dots, a_n$ , где  $a_i \in A$ ,  $A$  — множество аминокислот.
- $S = s_1, s_2, \dots, s_n$ , где  $s_i \in \{H, E, C\}$ :
  - $H = \alpha$  helix
  - $E = \beta$  sheets
  - $C = coil$ .

## Формализация задачи:

- $P = a_1, \dots, a_n$ , где  $a_i \in A$ ,  $A$  — множество аминокислот.
- $S = s_1, s_2, \dots, s_n$ , где  $s_i \in \{H, E, C\}$ :
  - $H = \alpha$  helix
  - $E = \beta$  sheets
  - $C = coil$ .
- **Input:**  $\mathbb{X} = \{P_1, \dots, P_n\}$ .
- **Output:**  $\mathbb{Y} = \{S_1, S_2, \dots, S_n\} \Rightarrow$  задача классификации.

# Типы вторичной структуры





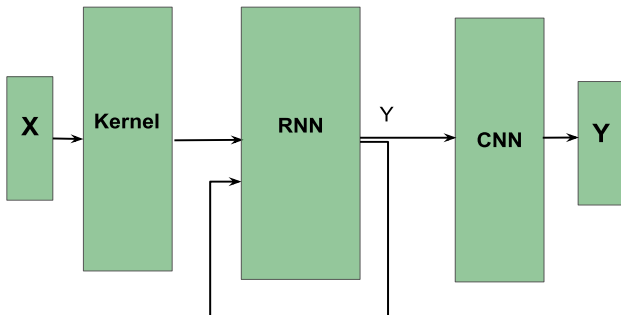
- 1 Сбор данных для обучения (проверка на репрезентативность).
- 2 Построение нейронной сети с использованием методов рекуррентных сетей.
- 3 Подбор параметров и характеристик (количество слоев, объем обучающей выборки).

- 1 Сбор данных для обучения (проверка на репрезентативность).
- 2 Построение нейронной сети с использованием методов рекуррентных сетей.
- 3 Подбор параметров и характеристик (количество слоев, объем обучающей выборки).
- 4 Обучение сети.
- 5 Проверка на тестовых данных.
- 6 Коррекция ошибок.

- **Рекуррентная сеть (RNN)** — нейронная сеть, в которой имеется обратная связь.

Функция активации: **softplus** =  $(1 + e^{-x})^{-1}$ .

- **Рекуррентная сеть (RNN)** — нейронная сеть, в которой имеется обратная связь.  
Функция активации: **softplus**  $= (1 + e^{-x})^{-1}$ .
- **Сверточная сеть (CNN)** — однонаправленная нейронная сеть, каждый входной сигнал подвергается нескольким слоям свертки.  
Функция активации: **softmax**.



## HYDROLASE 09-MAY-14 4D2I

### Chain: A



### Chain: B



Рис.: Визуализация вторичной структуры белка с помощью *Stride*.

## 1 Трудности

- Белки разной длины.
- Во вторичной структуре имеются и другие типы (не только  $\{H, E, C\}$ ).
- Обработка белков с более чем одной цепью.

## 2 Решения

- Взяли большую длину белка  $Inf = 400$ .
- Все типы вторичной структуры, не входящие в множество  $S$  кодируем как  $O$  — other.
- В данной работе обрабатывались белки только с одной цепью.

## 3 Данные

- Количество белков:  $n = 15000$ .
- Наилучшее предсказание: 74%.

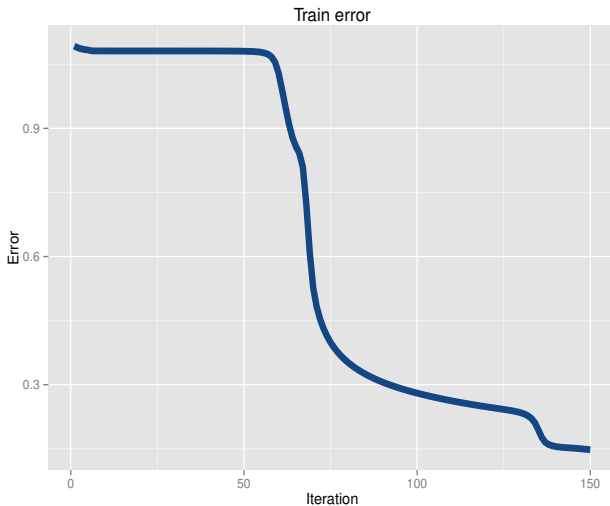


Рис.: Ошибка обучения нейронной сети на выборке train.

# Результат. Confusion matrix

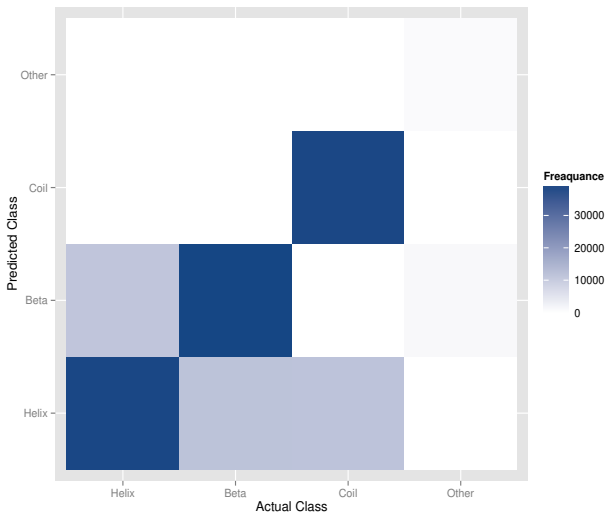


Рис.:  $n = 15000$ , совпадения 74%.