Предсказание вторичной структуры белка на основе методов Deep Learning

Караваева Валерия

Институт биоинформатики

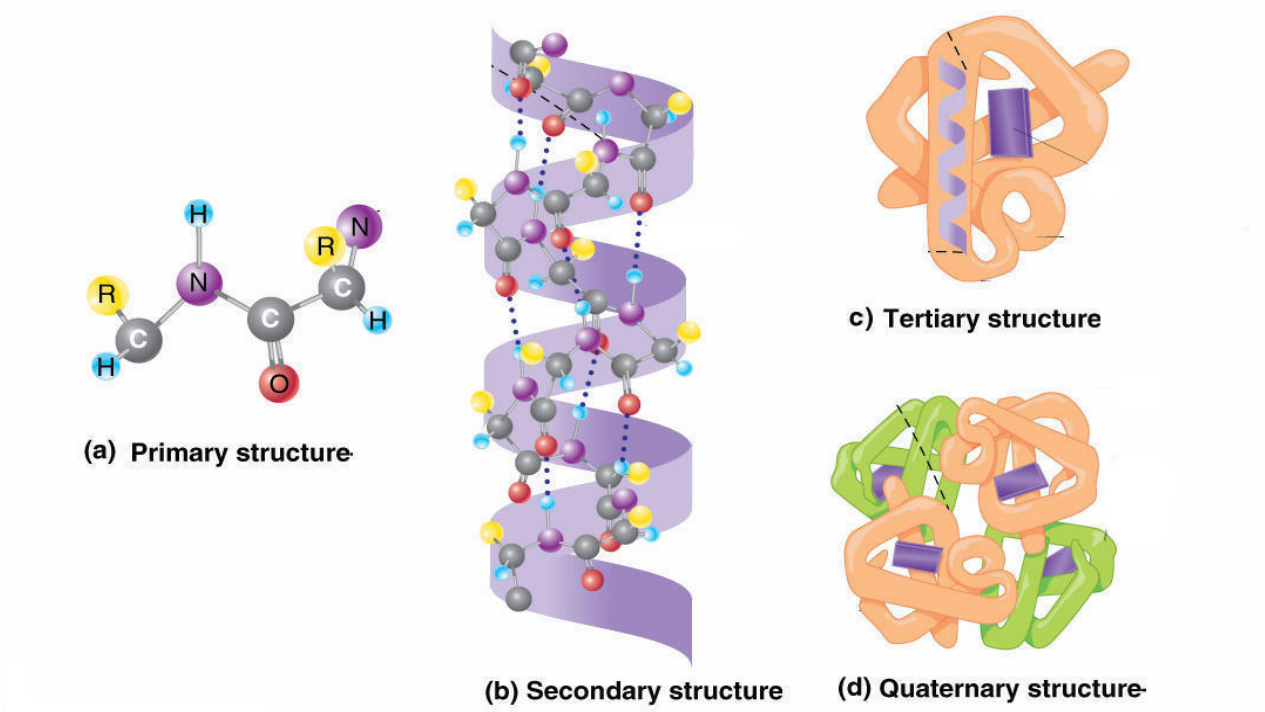
Руководитель — И. Дрокин

BIOCAD
Biopharmaceutical Company

Санкт-Петербург
2015г.

Мотивация изучения белков

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



Цели и задачи

Задача

Предсказание вторичной структуры белка с использованием первичной структуры.

Цели и задачи

Задача

Предсказание вторичной структуры белка с использованием первичной структуры.

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

Цели и задачи

Задача

Предсказание вторичной структуры белка с использованием первичной структуры.

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

Цели и задачи

Формализация задачи:

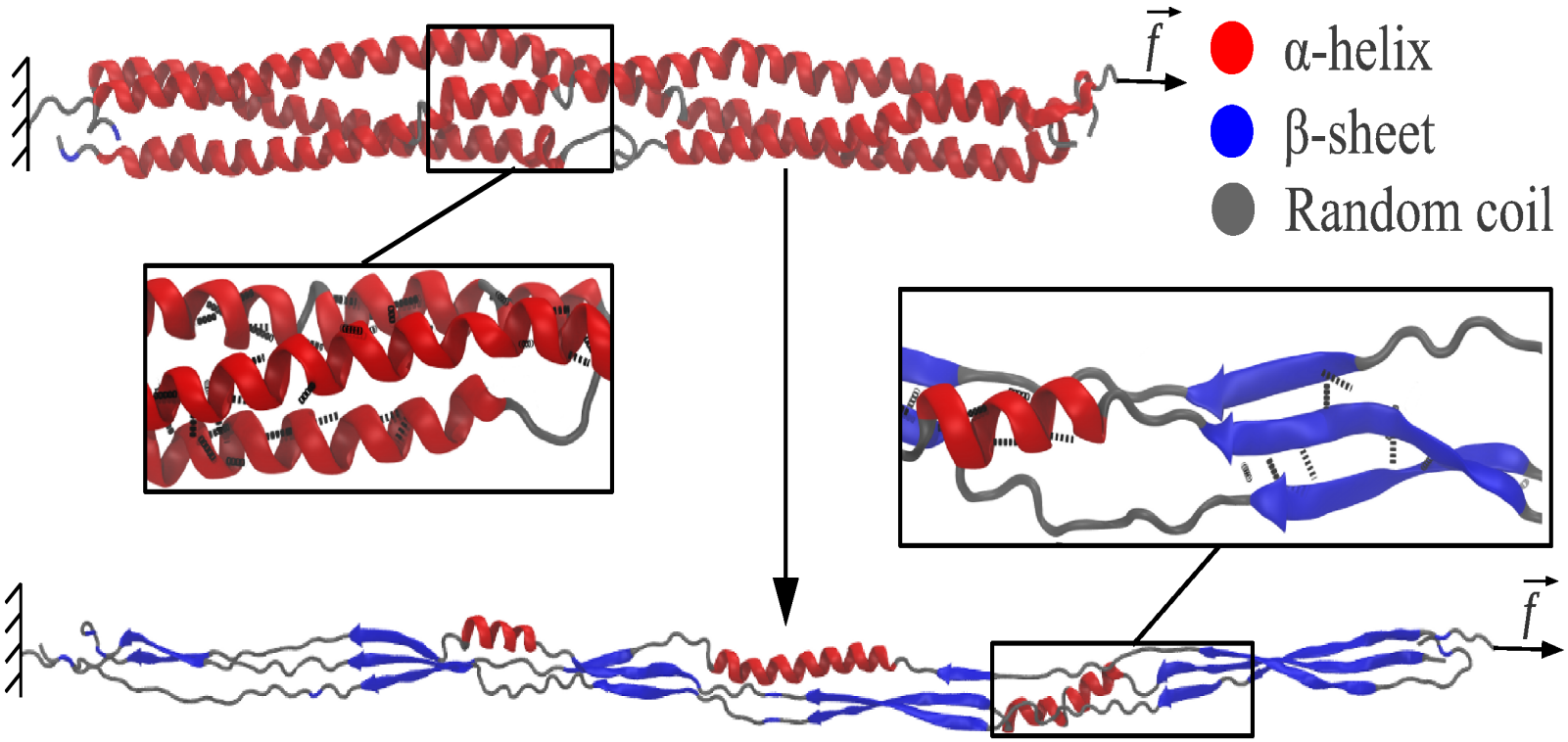
- $\mathrm{P} = a_1, \ldots, a_n$, *где* $a_i \in A$, A *— множество аминокислот.*
- $S = s_1, s_2, \ldots, s_n$, *где* $s_i \in \{\mathrm{H}, \mathrm{E}, \mathrm{C}\}$*:*
  - $\mathrm{H} = \alpha$ *helix*
  - $\mathrm{E} = \beta$ *sheets*
  - $\mathrm{C} =$ *coil.*

Цели и задачи

Формализация задачи:

- $\mathrm{P} = a_1, \ldots, a_n$, *где* $a_i \in A$, A *— множество аминокислот.*
- $S = s_1, s_2, \ldots, s_n$, *где* $s_i \in \{\mathrm{H}, \mathrm{E}, \mathrm{C}\}$*:*
  - $\mathrm{H} = \alpha$ *helix*
  - $\mathrm{E} = \beta$ *sheets*
  - $\mathrm{C} =$ *coil.*
- ***Input****:* $\mathbb{X} = \{\mathrm{P}_1, \ldots, \mathrm{P}_n\}$.
- ***Output****:* $\mathbb{Y} = \{S_1, S_2, \ldots, S_n\} \Rightarrow$ *задача классификации.*

Типы вторичной структуры



Этапы решения задачи

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).

Этапы решения задачи

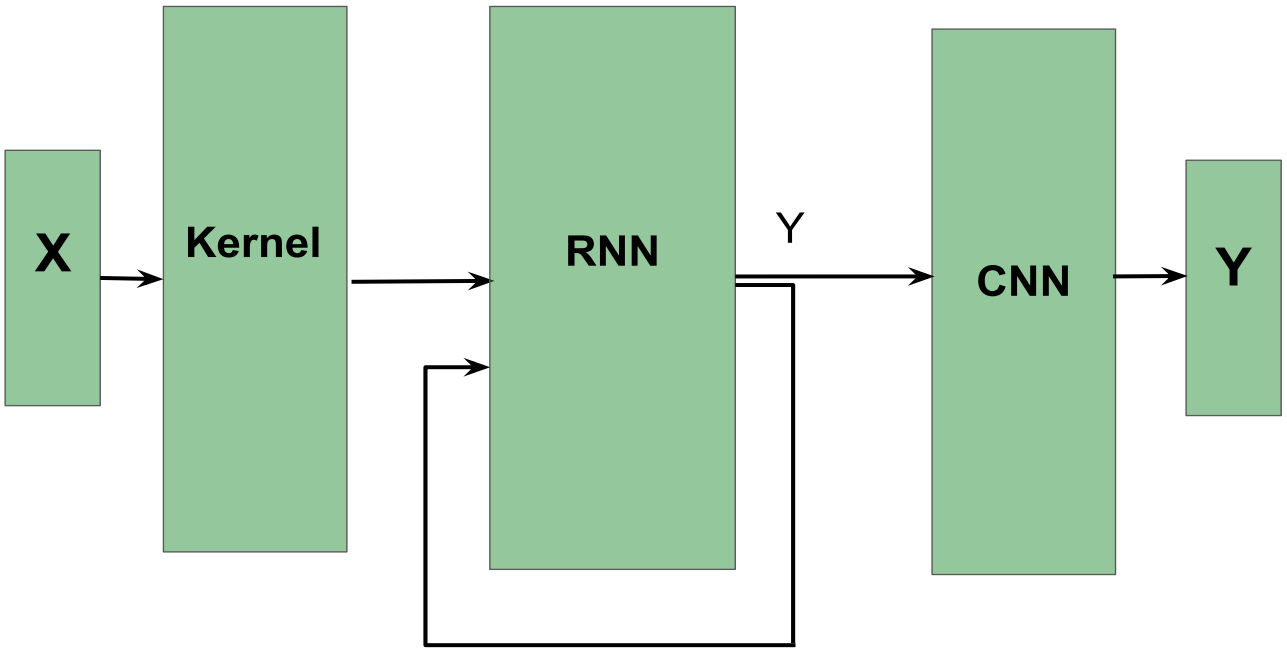
1. Сбор данных для обучения (проверка на репрезентативность).
2. Построение нейронной сети с использованием методов рекуррентных сетей.
3. Подбор параметров и характеристик (количество слоев, объем обучающей выборки).
4. Обучение сети.
5. Проверка на тестовых данных.
6. Коррекция ошибок.

Результат. Модель нейронной сети

- **Рекуррентная сеть (RNN)** — нейронная сеть, в которой имеется обратная связь.
  Функция активации: **softplus** $= (1 + e^{-x})^{-1}$.

Результат. Модель нейронной сети

- **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**.



Результат. Обучающая выборка. Stride

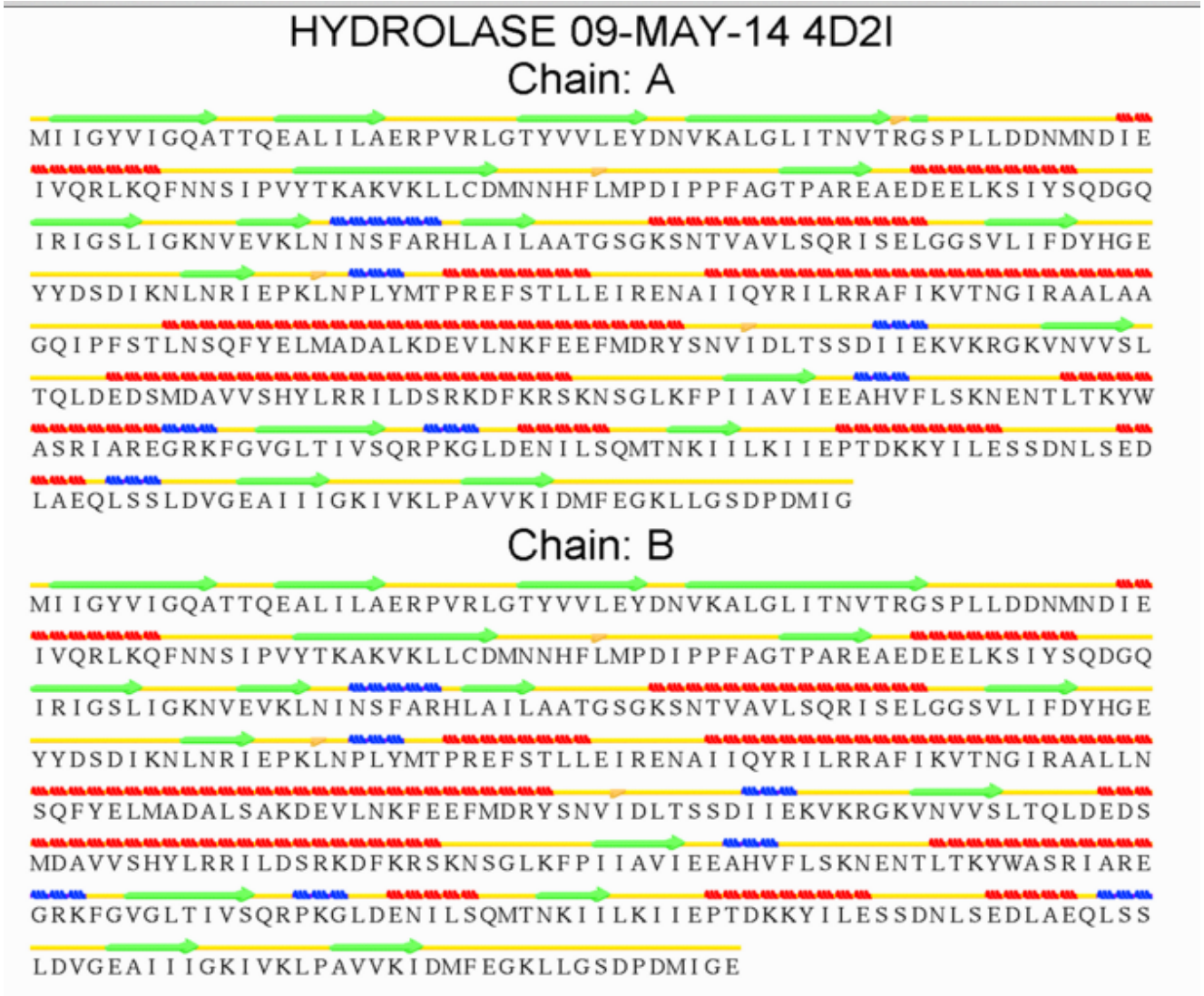


Рис.: Визуализация вторичной структуры белка с помощью *Stride*.

Результаты.

1. **Трудности**
   - Белки разной длины.
   - Во вторичной структуре имеются и другие типы (не только $\{\mathrm{H}, \mathrm{E}, \mathrm{C}\}$).
   - 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%.

Результат. Train error

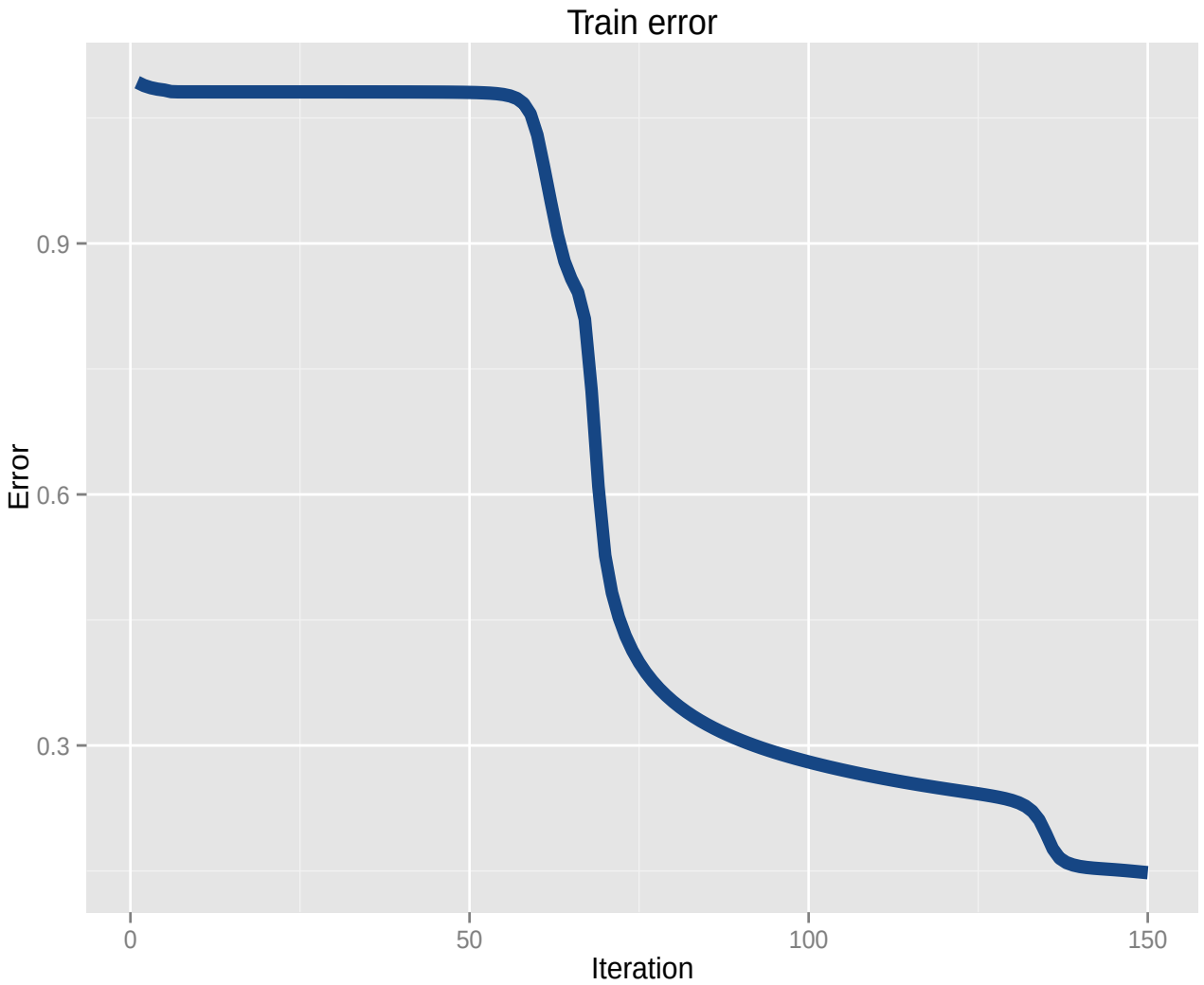


Рис.: Ошибка обучения нейронной сети на выборке train.

Результат. Confusion matrix

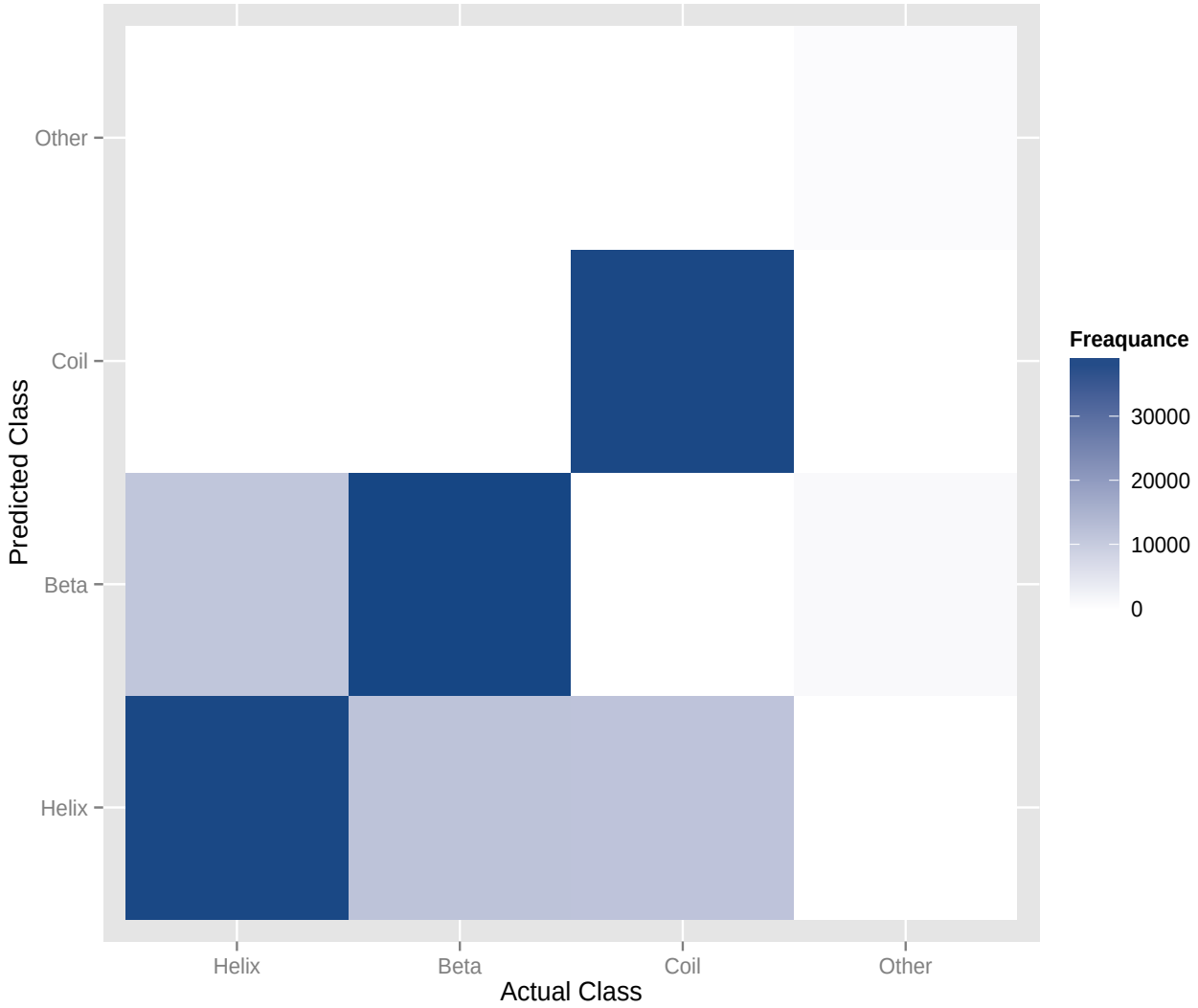


Рис.: n = 15000, совпадения 74%.