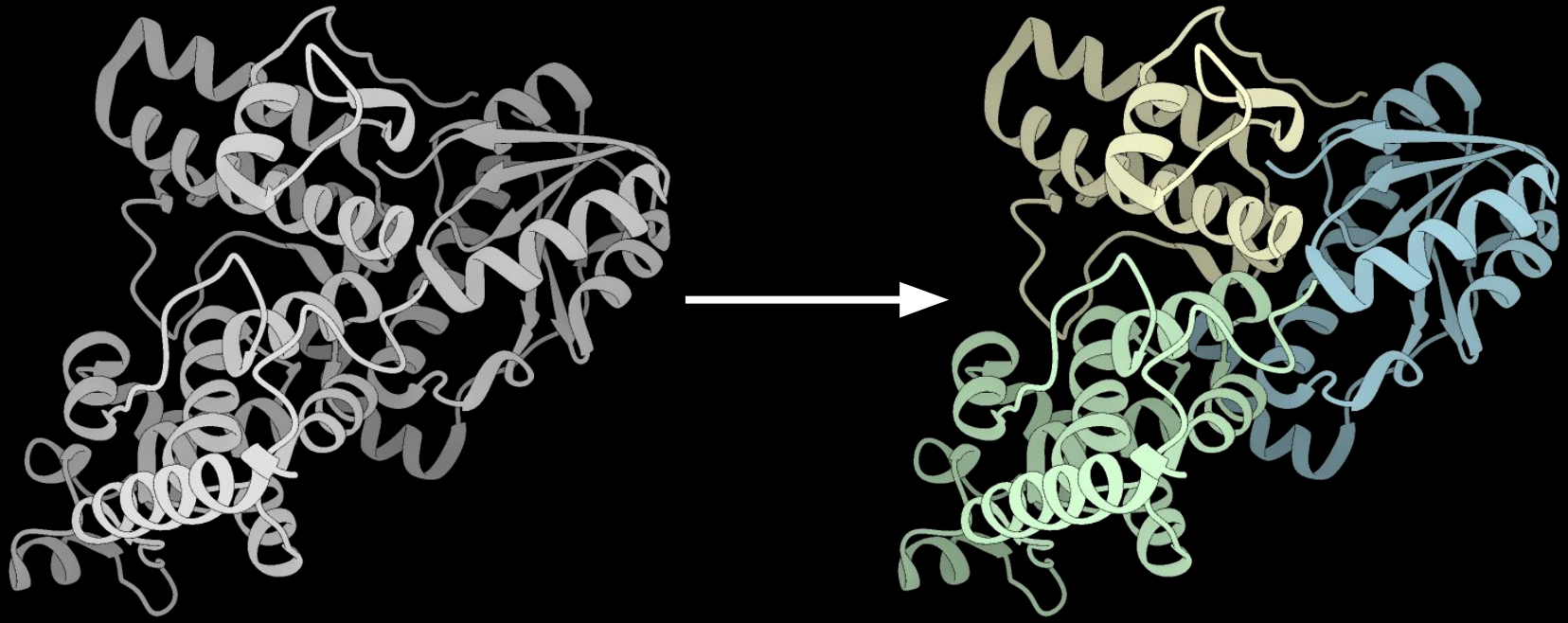
LifeCode

Поиск и классификация доменов в белках

Д. А. Яковлев, А. В. Кобченко

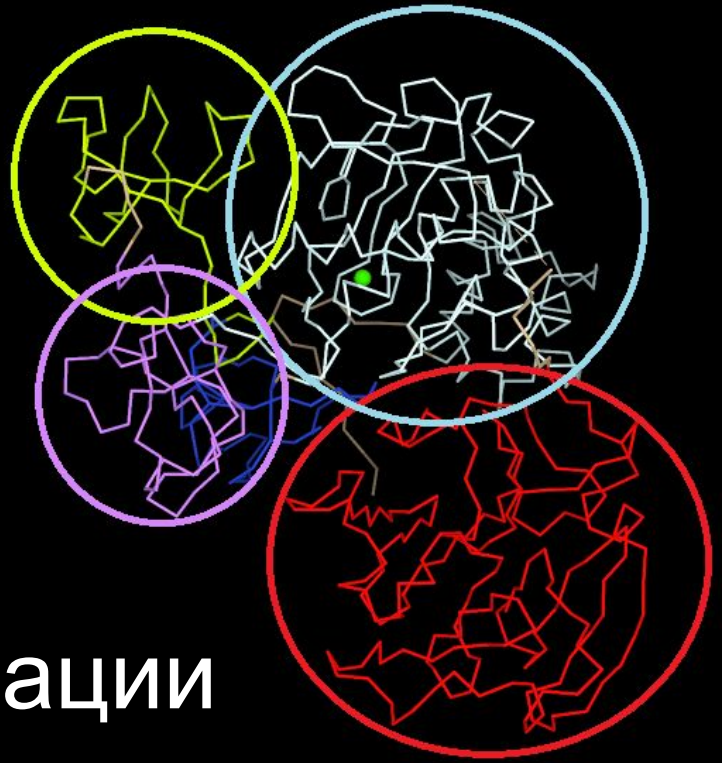
Задача — разметить домены в белке



Что такое домен

Домен — совокупность аминокислот, которые ближе друг к другу, чем ко всем остальным, *кластер*.

Предполагается, что каждый домен содержит гидрофобное ядро.



Задача о кластеризации

Уменьшение размерности данных

Все атомы из структуры не нужны — достаточно рассмотреть только C_α

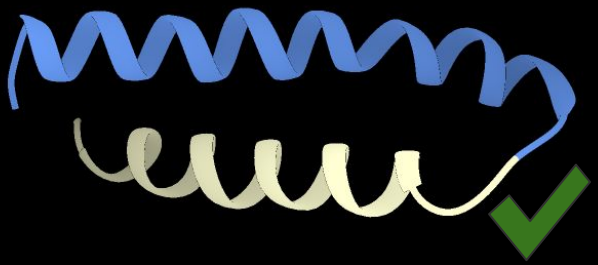
10^3–10^4 $\longrightarrow$ 10^2–10^3 атомов

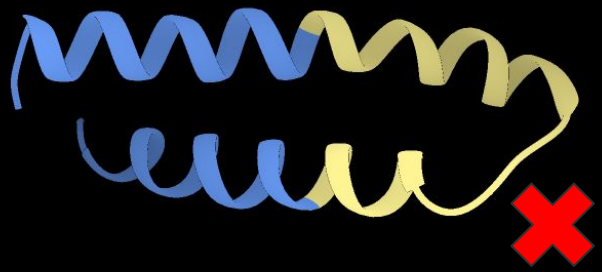
Вместо декартовых координат — матрица расстояний

$(x_i,y_i,z_i) \longrightarrow R_{ij}$

Генерация обучающей и тестовой выборок

Из базы данных белковых доменов CATH отбирались те записи, в которых границы доменов не находятся внутри элементов вторичной структуры

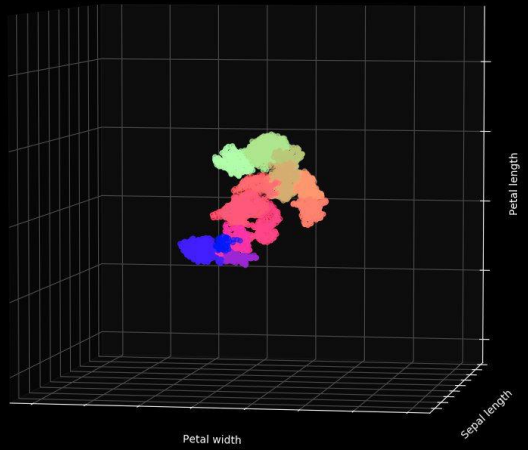




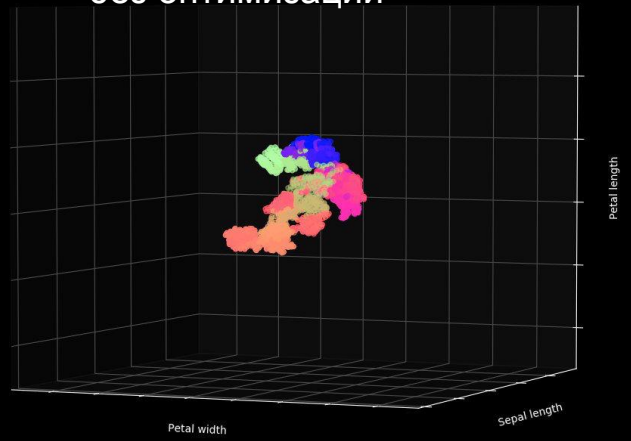
Кластеризация

BIRCH	< threshold
	< branching_factor

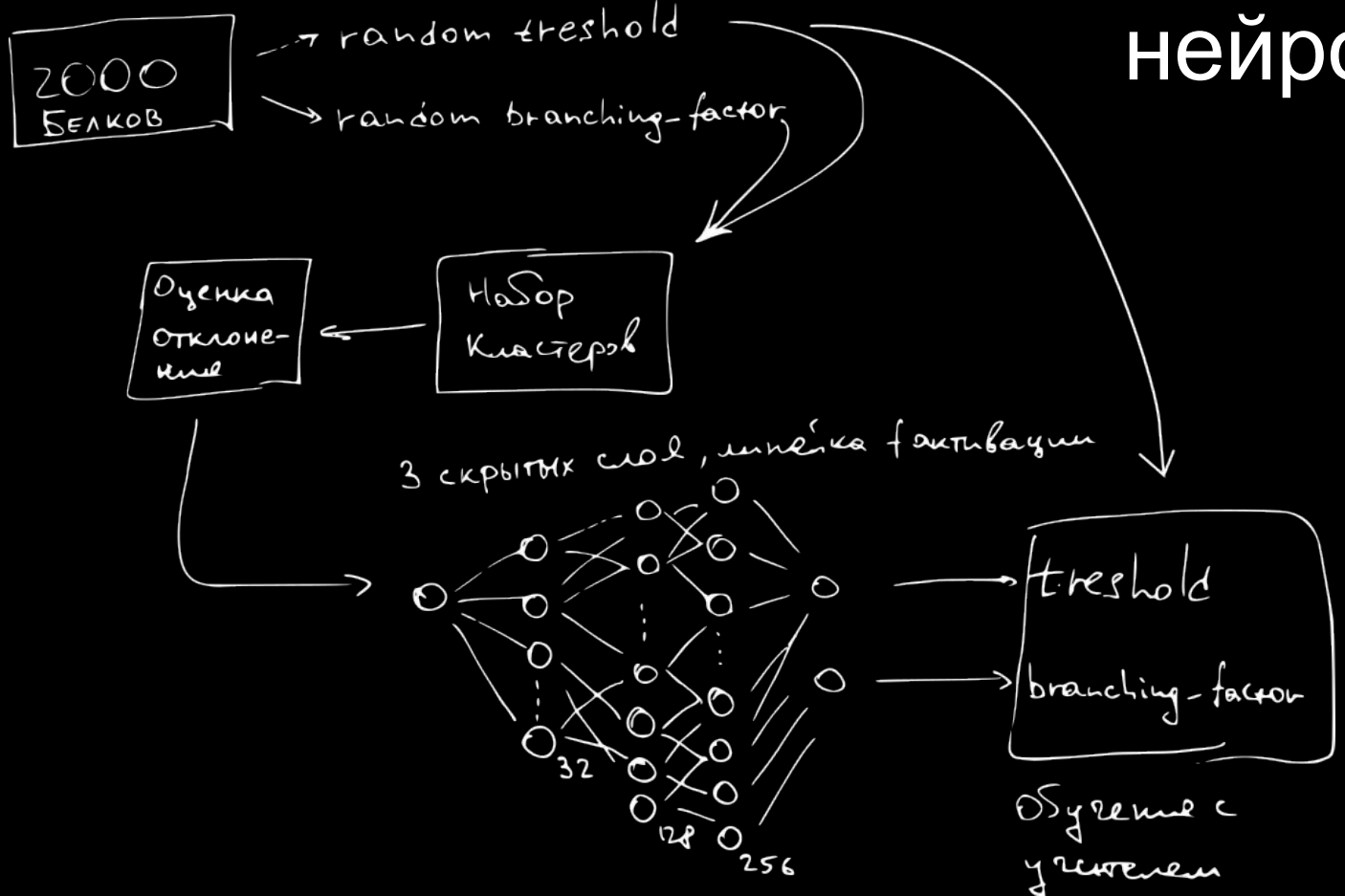
Как должно быть



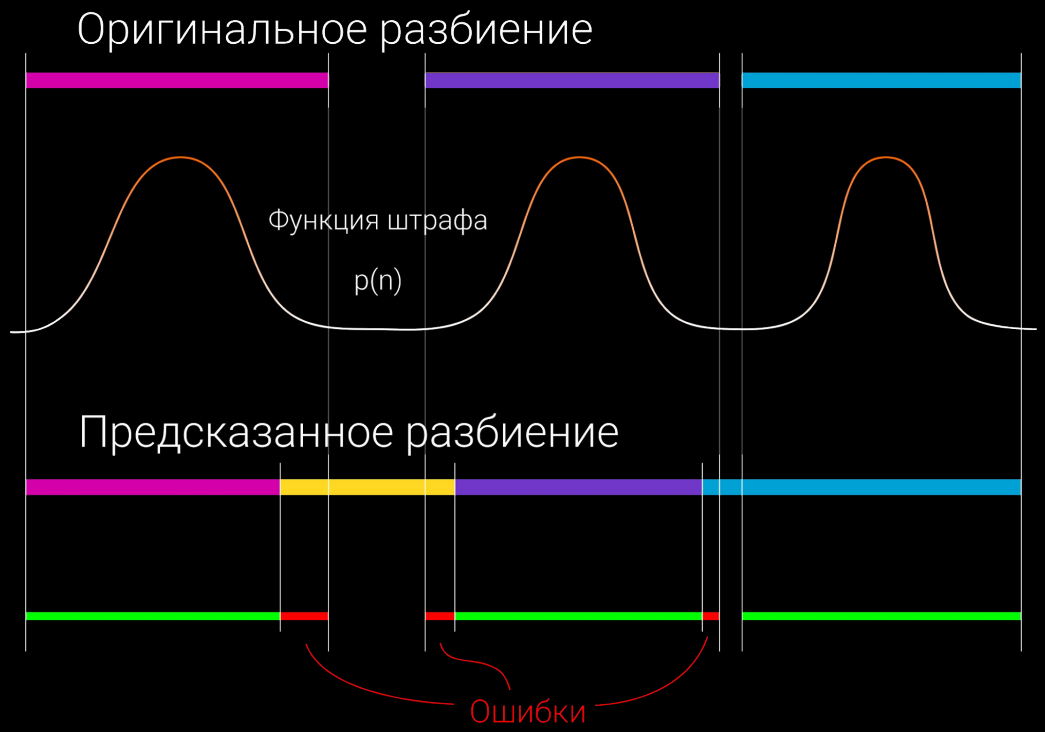
BIRCH
без оптимизации



Оптимизация параметров BIRCH нейросетью



Оценка отклонения



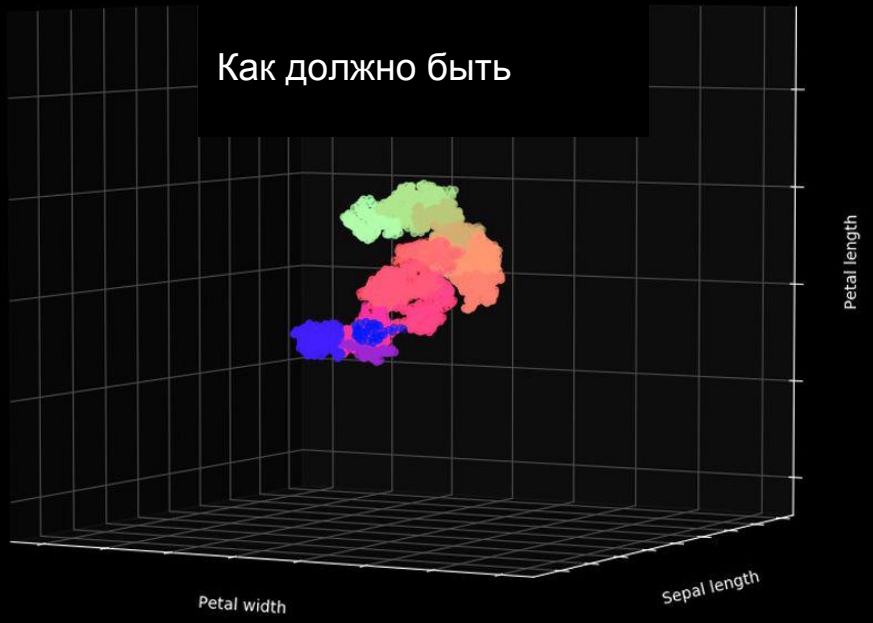
Виды штрафных функций:

- равномерная
- треугольная
- биномиальная

Результат

BIRCH, 33 000
итераций оптимизации

Как должно быть



Решенные проблемы

1. Предложена функция оценки разбиения (*scoring function*);
2. Значительно увеличена обучающая выборка (с 2 до 130 тыс. полипептидных цепей): примеры стали разнообразнее, сеть реже переобучается и можно использовать кросс-валидацию;
3. Добавлен ряд гиперпараметров, позволяющих тоньше настраивать алгоритм кластеризации.