

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Национальный исследовательский университет  
«Высшая школа экономики»**

Факультет компьютерных наук

Магистерская программа «Анализ данных в биологии и медицине»

Департамент анализа данных и искусственного интеллекта

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

На тему «Анализ связывания ионов магния с РНК»

Студент группы № мАДБМ16

Тихонова Полина

Руководитель КР

д.ф.-м.н., профессор,

базовая кафедра Яндекс

И.И. Цитович

Москва 2018

## Аннотация

Взаимодействие ионов магния с молекулами рибонуклеиновых кислот (РНК) в значительной степени определяет их структуру, которая в свою очередь определяет, какие функции будет выполнять РНК в клетках живых организмов. Это определяет актуальность изучения данных взаимодействий.

Существующие на данный момент экспериментальные методы определения структур РНК и их окружения (например, рентгеновская кристаллография) не позволяют с достаточной точностью идентифицировать сайты связывания ионов магния с РНК. Таким образом, является актуальной задача предсказания сайтов связывания с использованием вычислительных методов (например, с помощью решений уравнений биофизики или методами машинного обучения)

Целью данной работы является анализ существующих решений для задачи предсказания сайтов связывания РНК с лигандами, а также изучение возможностей применения современных методов машинного обучения к задаче предсказания участков РНК, взаимодействующих с ионами магния.

## Abstract

3D RNA structure is mainly determined by interactions between magnesium ions and RNA molecules and depending on the structure RNA plays different roles in living cells.

Nowadays, there are not any efficient experimental methods, that can identify binding sites between magnesium ions in RNA molecules. The existing methods are too complicated to be applied to big structures. Thus, the question of ion binding sites prediction is still opened and remains topical.

This problem can be solved by some computational and machine learning methods. The aim of this work is to find out efficiency of such methods and to analyze the existent ones.

# Оглавление

Аннотация	2
Abstract	2
1. Введение	4
1.1 Описание проблемы	4
1.2 Цель и структура работы	6
2. Анализ существующих алгоритмов	7
2.1 FEATURE	7
2.2 MetalionRNA	9
3. Общий вид задачи	10
3.1 Данные и формулировка задачи	10
3.2 Признаки	15
4. Модель	16
4.1 Выбор набора данных	16
4.2 Препроцессинг/ Обработка данных	23
4.2.1 Пропущенные значения	23
4.2.2 Анализ признаков	25
4.2.3 Анализ структур в выборке	28
4.3 Подбор параметров алгоритма	29
4.4 Алгоритм построения модели	31
4.5 Валидация модели	32
5. Результаты	34
5.1 Оценка работы текущих сервисов	34
5.2 Аппроксимация координат ионов магния	37
6. Заключение	44
7. Список литературы	45
Приложение А	46
Приложение В	47
Приложение С	48

# 1. Введение

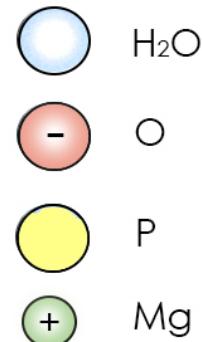
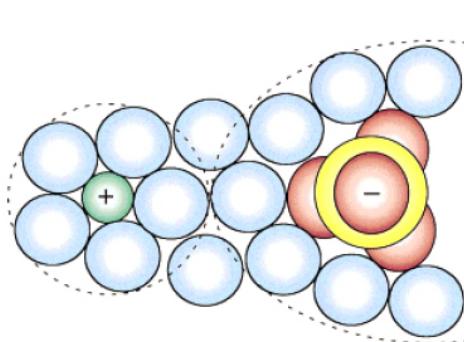
## 1.1 Описание проблемы

РНК, ДНК и белки являются основными биологическими макромолекулами в живых клетках. Однако, до недавнего времени наши представления о роли РНК в клетке ограничивались реализацией генетического кода. За последние два десятилетия было открыто большое количество типов некодирующих РНК, доля которых существенно превышает белок-кодирующие РНК [1]. Такие РНК могут выполнять транспортные, регуляторные и другие функции. А как известно [2], то, какие функции выполняет рибонуклеиновая кислота, напрямую зависит от ее пространственной структуры. Таким образом, важно изучать не только последовательность нуклеотидов в молекуле, но и ее расположение в пространстве.

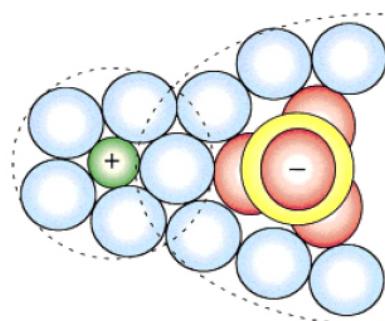
Поскольку сахарно-фосфатный остов молекулы РНК отрицательно заряжен, то большое влияние на формирование третичной структуры оказывают положительно-заряженные ионы металлов [3] [4]. Связи между ионами металлов и РНК образуются в водной среде, поэтому можно выделить 3 типа связей «ион металла – РНК» в зависимости от степени участия молекул воды в этом процессе [5]. Виды связей продемонстрированы на Рис. 1.1 Типы связываний на примере иона магния.

1. Диффузионное связывание. Полностью гидратированные катионы стабилизируют РНК. Ионы взаимодействуют с рибонуклеиновой кислотой на большом расстоянии, без прямых контактов с РНК.
2. Специфическое связывание через воду. Гидратированные катионы стабильно связываются с РНК через 1-2 слоя молекул воды.
3. Специфическое связывание с определенным сайтом на РНК. Частично гидратированные катионы взаимодействуют с РНК напрямую, создавая очень сильное взаимодействие.

1. Диффузионное связывание



2. Специфическое связывание через воду.



3. Специфическое связывание.

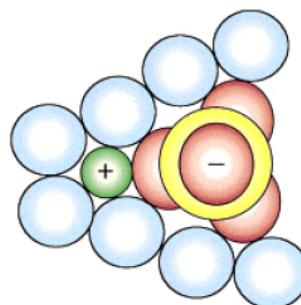


Рис. 1.1 Типы связываний

Наиболее распространенные металлы, которые образуют связи с рибонуклеиновыми кислотами – это натрий, калий и магний. Но самое большое влияние, а также самым часто встречающимся катионом, является магний [6] [7], так как из всех перечисленных металлов, ион магния имеет самый большой заряд на фоне самого маленького радиуса. Это позволяет ему плотнее организовывать молекулы воды вокруг себя, а высокий заряд является более эффективным в стабилизации пространственной структуры рибонуклеиновой кислоты. Таким образом, для формирования третичной структуры потребуется меньше ионов при более низкой энергии гидратации.

Рентгеновская кристаллография и ядерно-магнитный резонанс являются основными экспериментальными методами определения пространственных структур биомолекул. Однако зачастую электронные плотности ионов довольно трудно различить и некоторые сайты связывания на самом деле могут оказаться водой. Чтобы снизить число таких ошибок, эти эксперименты должны проводиться с очень высоким разрешением, что является довольно затратным. Поэтому на сегодняшний день достаточно актуальной задачей является разработка вычислительных методов, которые, если не заменят лабораторные исследования, то хотя бы помогут экспериментаторам более точно распознавать сайты связывания.

На самом деле такие методы существуют, например, Герман и Вестхов используют модель броуновской динамики диффузии ионов металлов для предсказания сайтов связывания [8], также для решения этой задачи могут использоваться и другие методы биофизики. Так, в работе [9] Мисра и Драпер определяли наиболее вероятные области расположения ионов с помощью уравнений Пуассона-Больцмана. Однако, эти методы являются очень затратными в вычислительном плане, поэтому также, как и экспериментальные технологии, они не позволяют работать с большими структурами, что существенно тормозит развитие науки в данной области.

По этой причине все еще актуален поиск методов, позволяющих наиболее точно определять сайты связывания ионов, которые при этом являлись бы наименее затратными и позволяли решать эту задачу на основании имеющихся данных.

Несмотря на кажущуюся невыполнимость задачи, аналогичная проблема была довольно успешно решена в протеомике: существуют многочисленные программы и сервисы, которые по заданной последовательности и/или структурной информации вычисляют сайты связывания ионов металлов с белками, при чем как распространенных, так и очень редких [10]. Но для задачи РНК такие алгоритмы почти отсутствуют.

Основными и единственными конкурентоспособными сервисами в данной области являются Feature [3] и MetallonRNA [11]. При этом Feature изначально существовал как онлайн-сервис для определения сайтов связывания ионов металлов с белками. Но позже его доработали и теперь он может решать аналогичную задачу для рибонуклеиновых кислот [3]. Однако эти алгоритмы не лишены недостатков: точность предсказания сильно зависит от конкретной структуры, так, в работе описаны результаты с точностью предсказаний в 96-97%, однако при валидации работы данных алгоритмов были получены далеко не впечатляющие результаты: например, для одной из структур алгоритм MetallonRNA смог распознать только 6 ионов магния из 24. Кроме того, сервис Feature на данный момент недоступен онлайн.

В следующей главе будет подробно разобран принцип работы каждого из алгоритмов, а в конце будет проведён сравнительный анализ на основе их предсказаний.

## 1.2 Цель и структура работы

---

Цель данной работы – на основе методов машинного обучения, собрать конкурентоспособный алгоритм, решающий задачу сайтов связывания ионов магния со структурами РНК на основе информации из базы данных URSDDB [12].

Структура работы:

- раздел 2 «Анализ существующих алгоритмов» посвящен изучению работы алгоритмов FEATURE и MetalionRNA;
- в разделе 3 «Постановка задачи и данные» подробно описаны данные, которые были предоставлены для решения поставленной задачи;
- в разделе 4 «Модель» пошагово выстраивается работа алгоритма и описываются проблемы, возникающие при работе с имеющимися данными;
- в разделе 5 «Результаты» проводится сравнение предложенного алгоритма с существующими сервисами, а также приводится аппроксимация координат ионов магния;
- раздел 7 «Выводы» завершает работу и подводит итоги;
- раздел 8 «Список литературы» - библиографический список;
- В раздел «Приложения» вошли дополнительные материалы.

## 2. Анализ существующих алгоритмов

### 2.1 FEATURE

С ростом числа экспериментально определенных пространственных структур РНК стало возможным применение статистического анализа при решении различных биологических задач, например, определения конкретных генов, мутации в которых приводят к различным заболеваниям. Задача распознавания сайтов связывания ионов с РНК не стала исключением.

Так, рассматриваемый алгоритм Feature сочетает в себе статистику и машинное обучение, чтобы предсказывать расположение ионов магния в пространственной структуре предоставленной рибонуклеиновой кислоты. Для осуществления этих предсказаний собираются все возможные биофизические свойства предполагаемых локаций, затем с помощью статистики выбираются те свойства, которые действительно могут свидетельствовать о наличии или отсутствие иона. И уже с помощью алгоритма машинного обучения на значениях этих признаков делается вывод о местонахождении иона магния.

Статистическая значимость биофизических признаков определяется с помощью U-критерия Манна — Уитни. При этом выявляется не только значимость конкретного признака, но и то, на каком удалении от предполагаемого сайта этот признак является значимым. Для этого берется набор элементов, про каждый из которых известно, является ли он сайтом или нет. Область вокруг каждого элемента делится на концентрические сферы и для каждого радиуса вычисляется рассматриваемый признак. Таким образом, получается  $n$  пар распределений (для сайтов и не сайтов), где  $n$  — число концентрических сфер. Эти распределения сравниваются с помощью U-критерия, который позволяет определить, является ли признак в данном слое значимым.

На Рис. 2.1 показан пример того, как вычисляется статистическая значимость для признака: количество атомов кислорода. Этот пример взят из статьи [3].

- A. Область вокруг каждого сайта и не сайта делится на концентрические окружности.
- B. Подсчитывается количество атомов кислорода, попавших в каждый слой.
- C. Составляются распределения. На картинке на этом этапе выписаны значения для данного свойства в третьем слое.
- D. Рассчитывается U-критерий Манна — Уитни, делается вывод о том, что распределения различны => Данное свойство в третьем слое является значимым.

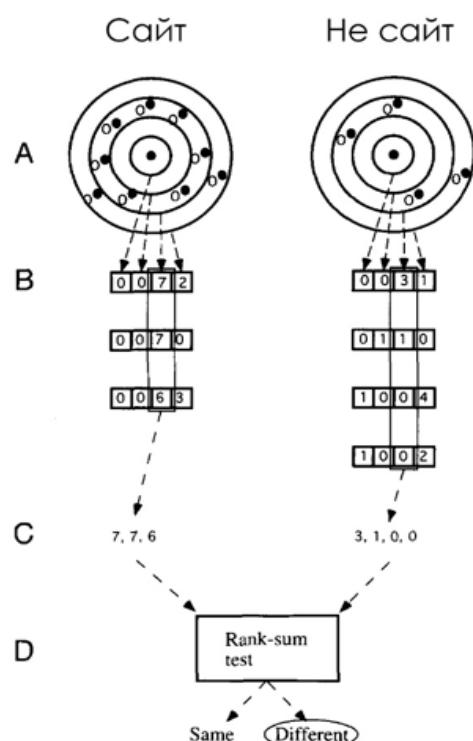


Рис. 2.1

Теперь, когда значимые признаки выбраны, можно работать с моделями машинного обучения. Feature в качестве такой модели использует байесовский классификатор с функцией оценки:

$$f_{\text{score}} = \sum_i \log \left( \frac{P(\text{Site} | v_i)}{P(\text{Not Site})} \right),$$

где  $v_i$  – признак, все признаки предполагаются независимыми.

Этот классификатор обучается на наборе данных: сайтов и не сайтов, для которых вычислены значения статистически значимых признаков, которые были определены на предыдущем шаге.

Блок-схема работы Feature приведена на Рис. 2.2.

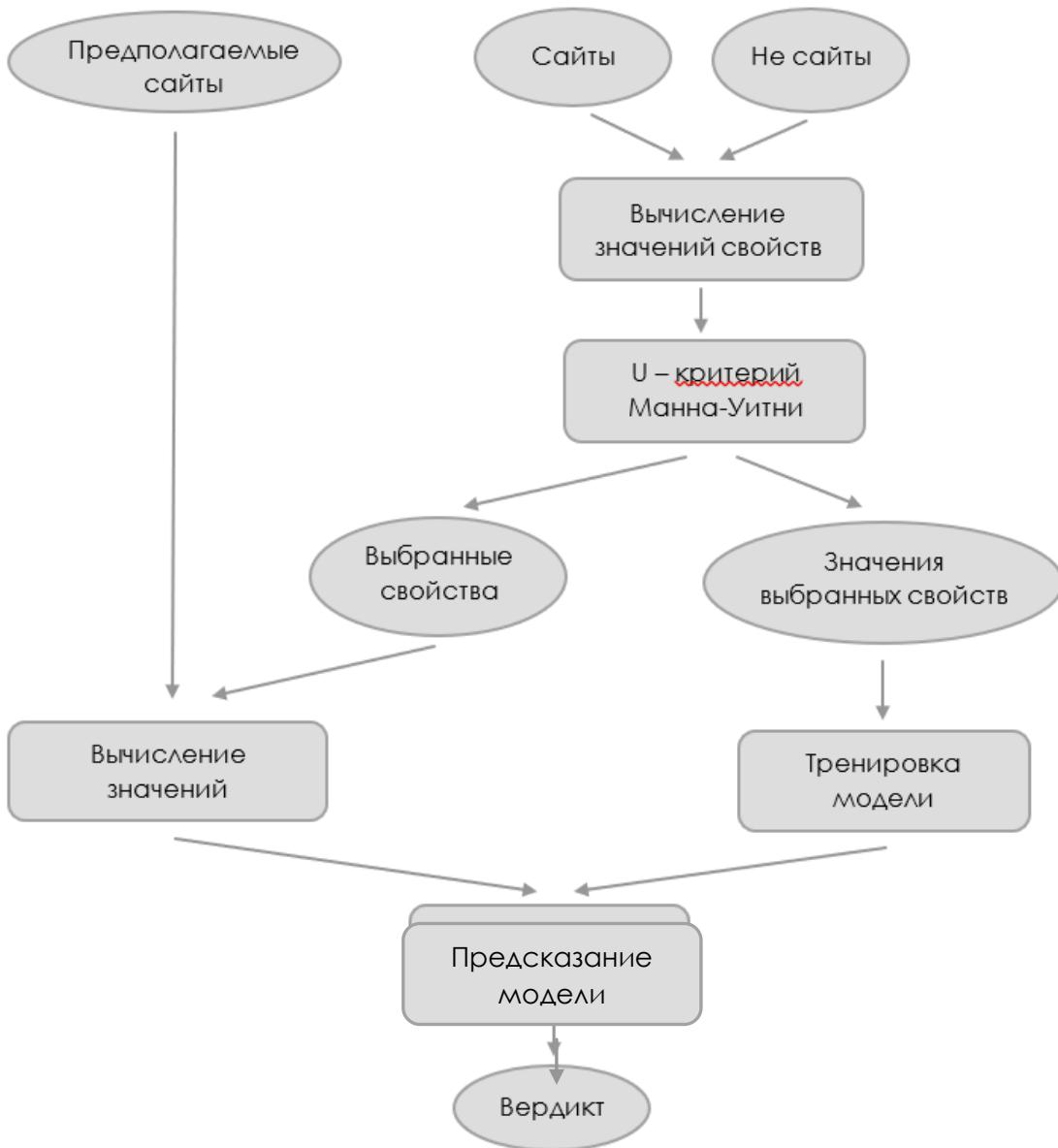


Рис. 2.2

## 2.2 MetallonRNA

Основная идея алгоритма MetallonRNA в том, что он анализирует окружение атомов РНК. Таким образом, для его работы требуется только третичная структура [11].

Более подробно: для каждой пары атомов РНК рассматривается сфера радиусом 9 Å вокруг одного из них (на Рис. 2.3 сфера нарисована вокруг атома b). Эта сфера делится на участки высотой 0.25 Å и шириной 5°. В каждом участке вычисляется анизотропный статистический потенциал

$$W(n)(d_1, \alpha_1; \dots, d_n, \alpha_n) = -RT \ln g^{(n)}(d_1, \alpha_1; \dots, d_n, \alpha_n),$$

где  $g^{(n)}$  - функция корреляции для n частиц, показывает экспериментально наблюдаемую частоту контактов катиона c со смежной парой атомов a, b; d - расстояние между ионом и атомом b;  $\alpha$  - угол ( $a, c, b$ ).

Этот потенциал помогает определить взаимодействия между белками и лигандами: участки, в которых вычисленный потенциал минимален, и есть предполагаемое местонахождение ионов.

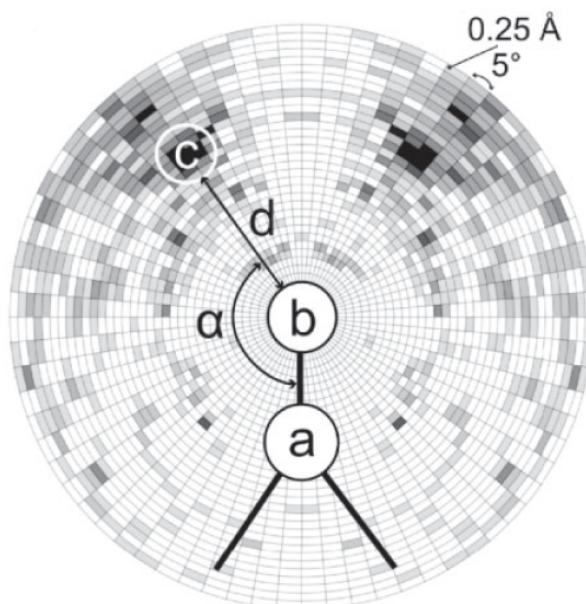


Рис. 2.3

Все значения (9 Å, 5° и 0.25 Å) подбирались опытным путем так, чтобы алгоритм определял ионы магния наиболее точно.

Забегая вперед, можно сказать, что MetallonRNA работает в большинстве случаев лучше, чем Feature. Наглядно это будет продемонстрировано в последней главе этой работы.

### 3. Общий вид задачи

#### 3.1 Данные и формулировка задачи

Все наборы данных в этой работе берутся из базы данных URSDB. Это аннотированная база пространственных структур РНК URSDB. В этой базе данных содержатся все РНК-содержащие документы из банка PDB [13]. Особенностью этой базы является то, что она позволяет скачивать сразу несколько цепочек в форматах mmCif или DSSR, а также предоставляет различную статистику по этим данным. В сумме в URS DataBase находится около 5000 структур, содержащих ионы металлов и большинство из этих ионов - магний.

Для данной работы из всех структур были отобраны только те, что имеют разрешение  $\leq 4 \text{ \AA}$ . Эти структуры разбиваются на классы эквивалентности [14], из которых впоследствии формируются 2 группы представителей: структуры с наилучшим разрешением и структуры с наибольшим числом ионов магния. В роли сайтов связывания могут быть как нуклеотиды, так и их фрагменты, или атомы, поэтому полученные группы рассматриваются с разной степенью подробности: на уровне нуклеотидов, фрагментов нуклеотидов или атомов. Также потенциальную связь элемента с ионом магния (прямую или через воду) мы будем рассматривать на разном удалении иона от элемента, а именно в радиусе 3  $\text{\AA}$ , 5  $\text{\AA}$  или 7  $\text{\AA}$ .

Для удобства, чтобы в последующем не описывать заново каждый набор данных словами ниже, в Таблица 3.1 **Ошибка! Источник ссылки не найден.** приведены названия всех 24 собранных выборок.

Таблица 3.1

	Нуклеотиды	Радиус	Фрагменты	Радиус	Атомы	Радиус
Наилучшее разрешение		3	minresol_N_	3	minresol_A_	3
		5		5		5
Макс. Число ионов		7		7		7
			maxmg_N_	3	maxmg_A_	3
				5		5
				7		7

Для каждого нуклеотида/фрагмента или атома в соответствующем наборе данных будет приведена информация о том, какой это нуклеотид/фрагмент/атом, какого типа спаривания он имеет, значения углов, вершиной которых он является и т.д. Подробнее эти признаки будут описаны в следующем разделе. Но главным(целевым) будет признак *tg*, который для каждого элемента показывает присутствие иона магния в окрестности соответствующего радиуса.

Таким образом, в этой работе мы будем стремиться предсказать, является ли объект (нуклеотид/фрагмент/атом) потенциальным сайтом связывания с ионом магния. Так как рассматриваются достаточно большие расстояния (до 7  $\text{\AA}$ ), то данные связи могут быть как прямыми, так и через воду.

В Таблица 3.2 – 3.7 приведены статистические данные обо всех 24 выборках: общее количество элементов в наборе, % и количество элементов, являющихся сайтами. В

Таблица 3.8 содержит справочная информация о строении фрагментов. Эти данные были собраны в ходе курсовой работы прошлого года [15].

Таблица 3.2 Статистика по нуклеотидам (наибольшее число ионов)

Нуклеотид	Количество нуклеотидов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расстоянии 3 - 7 Å
<b>Аденин</b>	22 595	2 055 (9 %)	6 544 (29 %)	9 302 (41 %)	9 302 (41 %)
<b>Цитозин</b>	18 997	1 161 (6 %)	4 603 (24 %)	7 820 (41 %)	7 820 (41 %)
<b>Гуанин</b>	23 963	2 387 (10 %)	8 409 (35 %)	10 872 (45 %)	10 872 (45 %)
<b>Урацил</b>	18 318	1 323 (7 %)	4 690 (26 %)	6 981 (38 %)	6 981 (38 %)
<b>Всего</b>	83 873	6 926 (8 %)	24 246 (29 %)	34 975 (42 %)	34 975 (42 %)

Таблица 3.3 Статистика по нуклеотидам (лучшее разрешение)

Нуклеотид	Количество нуклеотидов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расстоянии 3 - 7 Å
<b>Аденин</b>	22 595	1 662 (7 %)	5 526 (25 %)	8 113 (36 %)	8 113 (36 %)
<b>Цитозин</b>	18 997	881 (5 %)	3 727 (20 %)	6 752 (36 %)	6 752 (36 %)
<b>Гуанин</b>	23 963	1 770 (7 %)	6 985 (29 %)	9 375 (39 %)	9 375 (39 %)
<b>Урацил</b>	18 318	1 030 (6 %)	3 860 (21 %)	6 049 (33 %)	6 049 (33 %)
<b>Всего</b>	83 873	5 343 (6 %)	20 098 (24 %)	30 289 (36 %)	30 289 (36 %)

Таблица 3.4 Статистика по атомам (наибольшее число ионов)

	Фрагмент	Кол-во атомов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расстоянии 3 - 7 Å
O	<b>Основания</b>	79 650	1 378 (1.7 %)	8 139 (10 %)	15 306 (19 %)	14 682 (18 %)
	<b>Фосфаты</b>	335 720	5 736 (1.7 %)	34 654 (10 %)	72 420 (22 %)	69 660 (21 %)
	<b>Рибоза</b>	167 854	630 (0.4 %)	5 064 (3 %)	21 642 (13 %)	21 359 (13 %)
	<b>Всего</b>	583 224	7 744 (1 %)	47 857 (8 %)	109 368 (19 %)	105 701 (18 %)
N	<b>Основания</b>	326 623	819 (0.25 %)	17 006 (5 %)	60 530 (19 %)	60 162 (18 %)
	<b>Всего</b>	909 847	8 563 (0.9 %)	64 863 (7 %)	169 898 (19%)	162 879(18%)

Таблица 3.5 Статистика по атомам (лучшее разрешение)

	Фрагмент	Кол-во атомов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расст.
						3 - 7 Å
O	<b>Основания</b>	79 562	972 (1.2 %)	6 185 (7.8 %)	12 110 (15 %)	11 545 (15 %)
	<b>Фосфаты</b>	335 400	4 538 (1.4 %)	27 697 (8 %)	59 667 (18 %)	57 127 (17 %)
	<b>Рибоза</b>	167 694	303 (0.18 %)	3 237 (1.9 %)	16 697 (10 %)	16 524 (10 %)
	<b>Всего</b>	582 656	5 813 (1 %)	37 119 (6 %)	88 474 (15 %)	85 196 (15 %)
N	<b>Основания</b>	326 331	391 (0.11 %)	11 940 (4 %)	47 272 (15 %)	47 025 (14 %)
	<b>Всего</b>	908 987	6 204 (0.7 %)	49 059 (5 %)	135 746 (19%)	132 221(15%)

Таблица 3.6 Статистика по фрагментам (наибольшее число ионов)

Фрагмент	Количество фрагментов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расстоянии 3 - 7 Å
<b>Основания</b>	83 873	1 990 (2 %)	14 036 (17 %)	26 305 (31 %)	26 305 (31 %)
<b>Фосфаты</b>	83 873	5 016 (6 %)	16 473 (20 %)	26 533 (32 %)	26 529 (32 %)
<b>Рибоза</b>	83 873	611 (0.7 %)	4 184 (5 %)	14 724 (18 %)	14 721 (18 %)
<b>Всего</b>	251 619	7 617 (3 %)	34 693 (13 %)	67 562 (42 %)	67 555 (27 %)

Таблица 3.7 Статистика по фрагментам (лучшее разрешение)

Фрагмент	Количество фрагментов	Связанных с Mg <sup>2+</sup> на расстоянии 3 Å	На расстоянии 5 Å	На расстоянии 7 Å	На расстоянии 3 - 7 Å
<b>Основания</b>	83 873	1 295 (2 %)	10 889 (13 %)	21 787 (26 %)	21 787 (26 %)
<b>Фосфаты</b>	83 873	4 102 (5 %)	13 542 (16 %)	22 506 (27 %)	22 505 (27 %)
<b>Рибоза</b>	83 873	297 (0.35 %)	2 860 (3 %)	11 705 (14 %)	11 704 (14 %)
<b>Всего</b>	251 619	5 694 (2 %)	27 291 (11 %)	55 998 (22 %)	55 996 (22 %)

Таблица 3.8 Сведения о строении фрагментов

Фрагмент	Количество «тяжелых атомов»	Количество N	Количество O	Количество N, O
<b>Основания</b>	<b>A</b>	10	5	0
	<b>C</b>	8	3	1
	<b>G</b>	11	5	1
	<b>U</b>	8	2	2
<b>Фосфаты</b>	5	0	4	4
<b>Рибоза</b>	7	0	2	2

Далее приведены гистограммы распределения элементов вокруг ионов магния на расстоянии 5 и 7 Å (Рис. 3.2-3.2) и наоборот, гистограммы расстояний от ионов магния до элементов (Рис. 3.3).

В среднем, расстояние от элемента до иона магния при сайт-специфическом связывании составляет  $\approx 2.1 \text{ \AA}$ , а при связывании через воду  $\approx 4 \text{ \AA}$  [8]. Тогда на основе этих гистограмм можно сказать, что

- Так как пик гистограммы расстояний от ионов до оснований приходится на 4 Å, то большинство сайтов среди оснований связаны с ионом магния через воду.
- Аналогично для рибозы. Только здесь пик гистограммы приходится на 6 Å, то есть большинство сайтов среди рибозы связаны с ионом магния через 2 молекулы воды.
- Однако в расстояниях от ионов до фосфатов имеется 2 значительных пика: 2.3 Å и 4 Å. То есть фосфаты склонны связываться с ионами магния как через воду, так и напрямую.

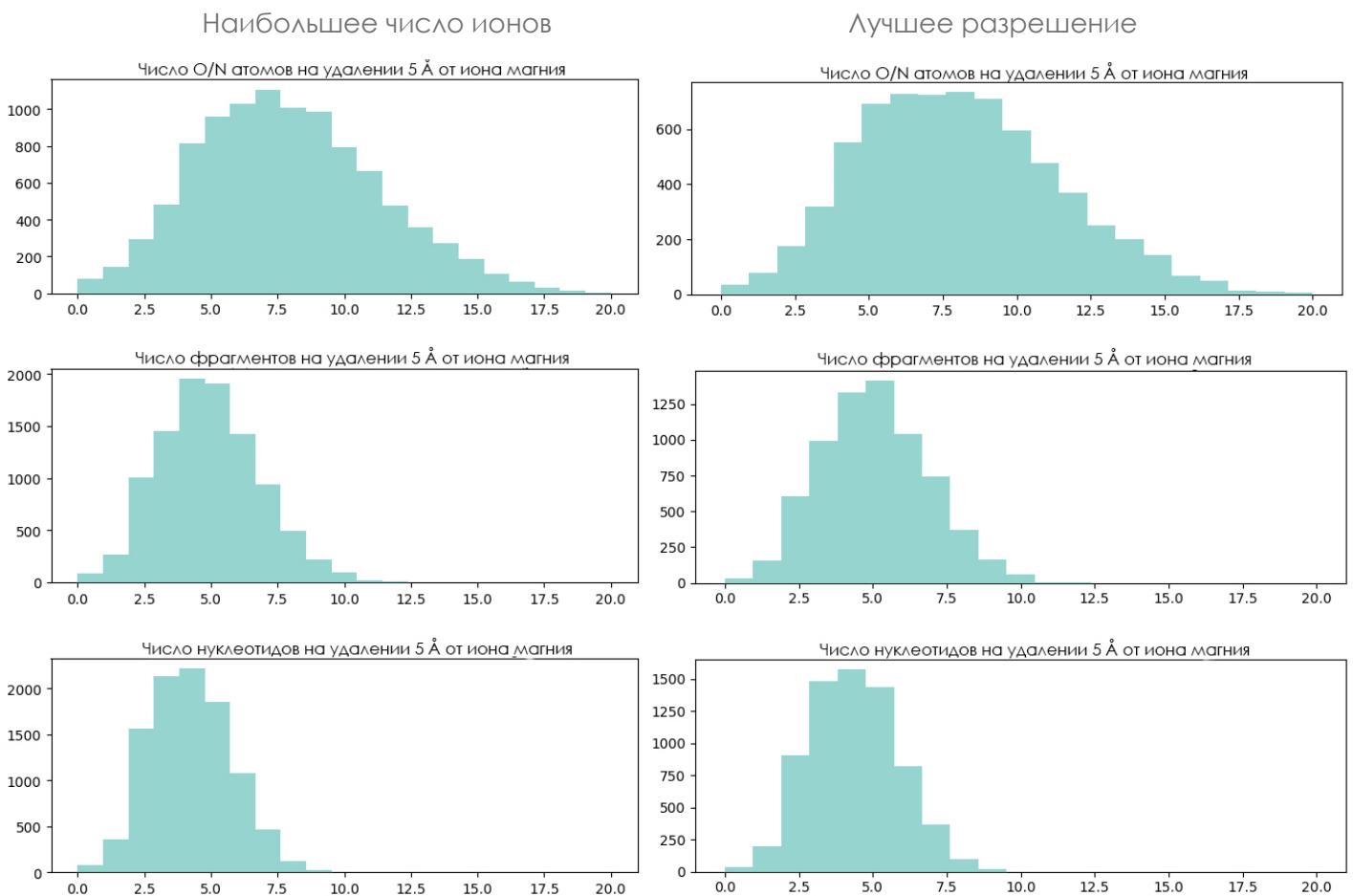


Рис. 3.2 Гистограммы распределения элементов вокруг ионов магния на расстоянии 5 Å.

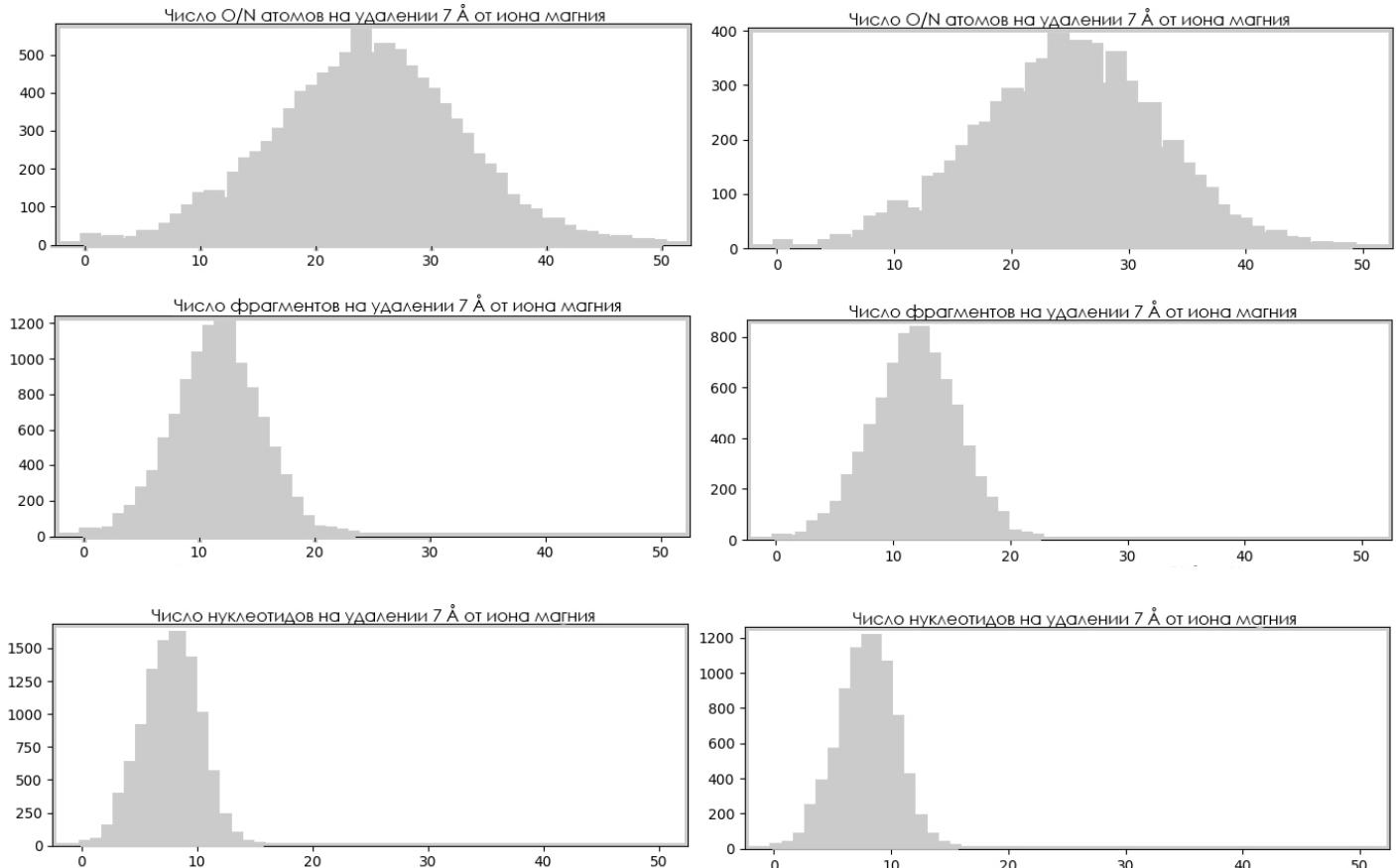
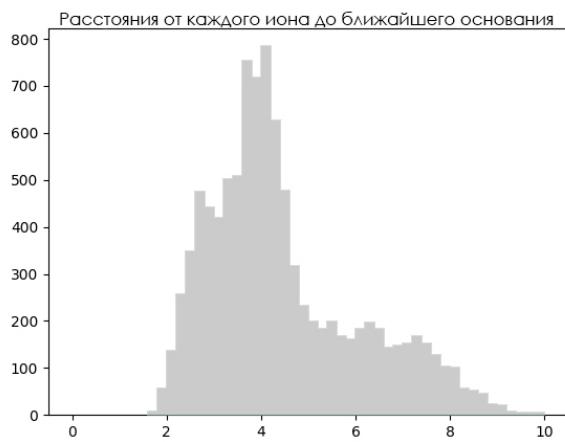


Рис. 3.1 Гистограммы распределения элементов вокруг ионов магния на расстоянии 7 Å.

Наибольшее число ионов



Лучшее разрешение

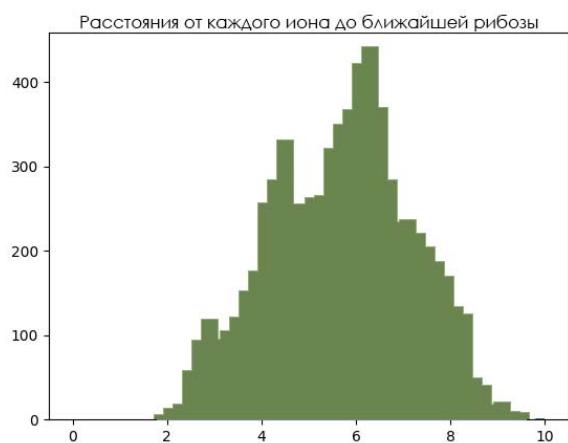
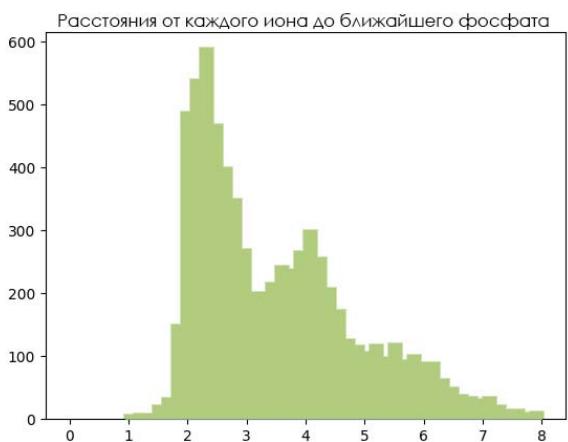
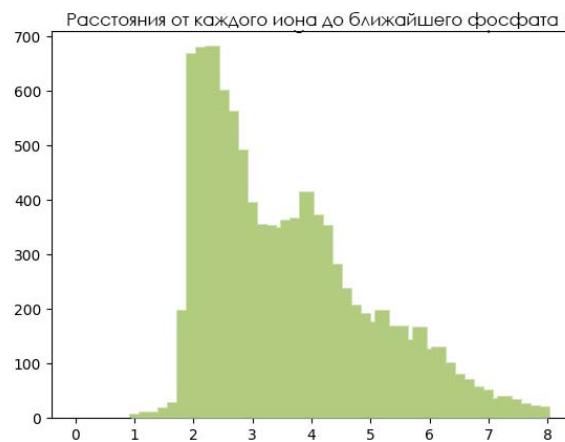
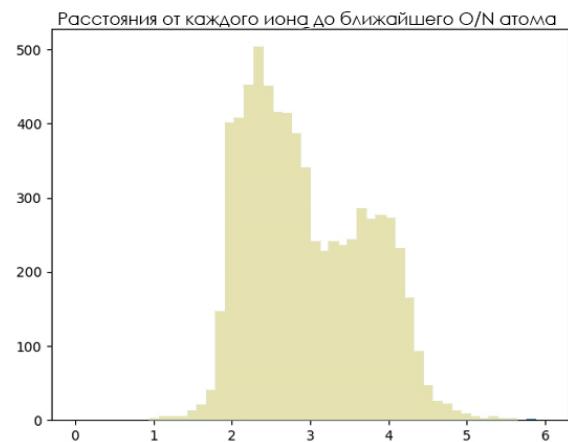
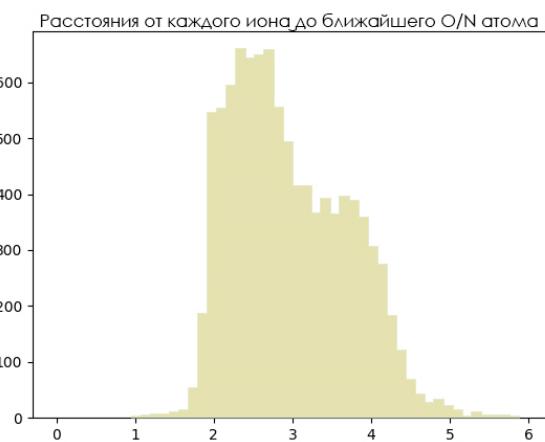
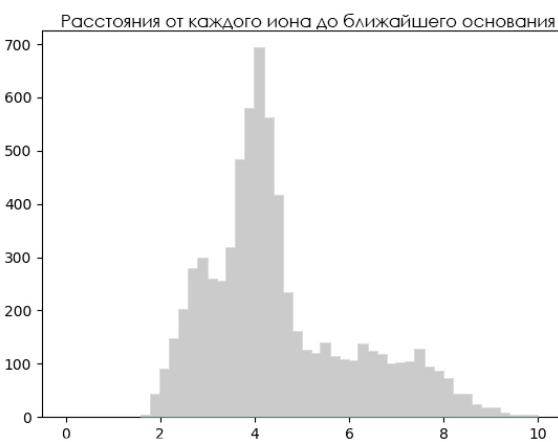


Рис. 3.3 Гистограммы расстояний от ионов до элементов.

## 3.2 Признаки

Для каждого элемента описано около 360 признаков. Их можно разделить на 5 групп, общих для всех типов сайтов – нуклеотидов, атомов, фрагментов нуклеотидов.

Данная информация также взята из анализа прошлого года [15].

<b>1. Общие данные о структуре.</b>	<b>4</b>
<ul style="list-style-type: none"><li>Файл получен рентгеновской кристаллографией или нет (<i>xray</i>).</li><li>Разрешение структуры (<i>resol</i>).</li><li>Длина цепи РНК (<i>chainlen</i>).</li><li>В файле присутствует белок (<i>protein</i>).</li></ul>	
<b>2. Данные о текущем нуклеотиде.</b>	<b>70</b>
<ul style="list-style-type: none"><li>Какой нуклеотид (<i>seqA</i>, <i>seqC</i>, <i>seqG</i>, <i>seqU</i>).</li><li>Торсионные углы нуклеотида (<i>alpha</i>, <i>beta</i>, <i>gamma</i>, <i>delta</i>, <i>epsilon</i>, <i>zeta</i>, <i>e-z</i>, <i>chi</i>, <i>phase-angle</i>, <i>sszp</i>, <i>dp</i>, <i>splay</i>, <i>eta</i>, <i>theta</i>, <i>etap</i>, <i>thetap</i>, <i>etapp</i>, <i>thetapp</i>, <i>v0</i>, <i>v1</i>, <i>v2</i>, <i>v3</i>, <i>v4</i>, <i>tm</i>, <i>p</i>). (см. Приложение А)</li><li>Спаривания (по классам) (<i>SI</i>, <i>SII</i>, <i>SIII</i>, ..., <i>SXXIX</i>, <i>cWH</i>, <i>cWS</i>, <i>cWW</i>, <i>cHS</i>, <i>cHH</i>, <i>cSS</i>, <i>tWH</i>, <i>tWS</i>, <i>tWW</i>, <i>tHS</i>, <i>tHH</i>, <i>tSS</i>). (см. Приложения В, С)</li></ul>	
<b>3. Тоже данные о соседних нуклеотидах – по 2 вверх и вниз по цепи.</b>	
<ul style="list-style-type: none"><li>К названиям признаков добавляется <math>1/2/m1/m2</math> (<math>m = minus</math>).</li></ul>	<b>280</b>
<b>4. Данные о вторичной структуре, которой принадлежит нуклеотид.</b>	<b>7</b>
<ul style="list-style-type: none"><li>Принадлежит ли нуклеотид стему (<i>stem</i>).</li><li>Тип петли (<i>hairpin</i>, <i>bulge</i>, <i>internal</i>, <i>junction</i>).</li><li>Длина нити/крыла (<i>wtlen</i>).</li><li>Номер нуклеотида в нити/крыле (мин. из двух нумераций) (<i>wtnum</i>).</li></ul>	
<b>5. Связан ли с <math>Mg^{2+}</math>. (<i>mg</i>)</b>	<b>1</b>

Признак, присутствующий в файлах, содержащих данные о фрагментах нуклеотидов и атомах:

<b>6. Информация о том, какой это фрагмент: фосфат, сахар или основание.</b>	<b>1</b>
--	----------

Признак, присутствующий в файлах, содержащих данные об атомах:

<b>7. Информация о атоме.</b>	<b>19</b>
<ul style="list-style-type: none"><li>Кислород или азот.</li><li>Имя атома (какой именно это кислород или азот).</li></ul>	

### Итого:

Для нуклеотидов описан 361 признак + 1 целевой.

Для фрагментов – 364 признака + 1 целевой.

Для атомов – 383 признака + 1 целевой.

## 4. Модель

### 4.1 Выбор набора данных

В качестве основного классификатора был выбран RandomForestClassifier. Так как именно этот классификатор наименее склонен к переобучению, а также он позволяет работать с признаками разных типов одновременно.

Суть алгоритма:

- из обучающей выборки генерируется подвыборка с повторениями размером, равному обучающей выборке;
- по этой выборке строится решающее дерево;
- эти две операции проделываются  $n$  раз, таким образом мы получаем  $n$  решающих деревьев;
- выбор класса происходит голосованием – выбор большинства.

Чтобы выяснить, как ведет себя RandomForest на разных выборках, была проведена кросс валидация алгоритма на каждом наборе.

Кросс валидация:

- в предоставленных наборах около 5% строк содержат пропущенные значения, на данном этапе эти строки будут опущены;
- каждый набор делится случайным образом на тренировочную и тестовую выборки, при чем тестовая составляет около 30% от исходной выборки;
- для каждой пары выборок алгоритм RandomForest с параметрами, установленными по умолчанию, тренируется на первой выборке и делает предсказания на второй;
- для оценки качества предсказания вычисляется его точность: отношение числа элементов с правильно предсказанным целевым признаком к общему числу элементов;
- процедуры деления набора на выборки, тренировки алгоритма, предсказания и вычисления точности повторяется по 15 раз для каждого набора;
- таким образом, чтобы получить общую оценку работы алгоритма на наборе данных, точности, полученные на соответствующих 15 разбиениях, усредняются.

Ниже, в Таблица 4.1-4.2 приведены данные кросс валидации.

Таблица 4.1 Макс. число ионов

Элементы	Радиус	%сайтов	Точность
Нуклеотиды	3 Å	8 %	0.92
Нуклеотиды	5 Å	29 %	0.76
Нуклеотиды	7 Å	42 %	0.7
Нуклеотиды	3-7 Å	41 %	0.71
Фрагменты	3 Å	3 %	0.97
Фрагменты	5 Å	14 %	0.87
Фрагменты	7 Å	27 %	0.81
Фрагменты	3-7 Å	27 %	0.87
Атомы	3 Å	0.9 %	0.97
Атомы	5 Å	7 %	0.91
Атомы	7 Å	19 %	0.83
Атомы	3-7 Å	18 %	0.83

Таблица 4.2 Лучшее разрешение

Элементы	Радиус	%сайтов	Точность
Нуклеотиды	3 Å	6 %	0.94
Нуклеотиды	5 Å	24 %	0.78
Нуклеотиды	7 Å	36 %	0.7
Нуклеотиды	3-7 Å	36 %	0.7
Фрагменты	3 Å	2 %	0.97
Фрагменты	5 Å	11 %	0.9
Фрагменты	7 Å	22 %	0.83
Фрагменты	3-7 Å	21 %	0.82
Атомы	3 Å	0.7 %	0.95
Атомы	5 Å	5 %	0.96
Атомы	7 Å	15 %	0.81
Атомы	3-7 Å	15 %	0.82

Согласно этим таблицам можно выделить наборы данных, на которых даже ненастроенный алгоритм дает высокий результат – точность свыше 90%. Однако, можно заметить явную зависимость между долей сайтов в наборе и точностью предсказания. Зная долю сайтов в выборке, можно предугадать точность предсказания, которую покажет алгоритм RandomForest: accuracy  $\approx$  1- доля сайтов.

Чтобы выявить причину этого феномена для тестовой выборки последнего разбиения набора нуклеотидов с наилучшим разрешением, связанных с ионами магния на расстоянии до 5 Å, была построена матрица неточностей (confusion matrix) (Рис. 4.1). Эта матрица наглядно показывает, что алгоритм RandomForest так и не научился распознавать наличие связи между элементом, в данном случае – нуклеотидом, и ионом магния. Наоборот, ввиду того, что количество не связанных с магнием нуклеотидов существенно превышает количество нуклеотидов, связанных с магнием, алгоритм посчитал более выгодной стратегию, при которой он по умолчанию считает все элементы несвязанными с магнием и только в каких-то особых случаях разрешает наличие связи.

- Элементы выборки – нуклеотиды;
- расстояние от элемента до  $Mg^{2+}$  - до 5 Å;
- разрешение – наилучшее;
- тестовая выборка последнего разбиения кросс валидации;
- % сайтов связывания – 29%.

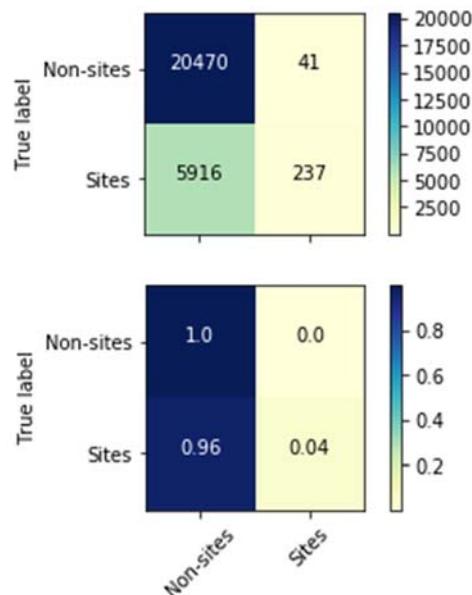


Рис. 4.1

Таким образом, становится понятно, что тренировать RandomForest на несбалансированной выборке в корне неверно, а значит необходимо преобразовать тренировочную выборку так, чтобы количества связанных и несвязанных с магнием элементов были одинаковыми. Это можно сделать двумя способами:

1. Количество сайтов равно исходному количеству несайтов. В таком наборе несайты – все элементы тренировочной выборки, несвязанные с магнием, сайты – случайная выборка с повторениями из элементов, имеющих связь с ионами магния, размер равен количеству несайтов.  
Общий размер тренировочной выборки увеличивается.
2. Количество несайтов равно исходному количеству сайтов. В таком наборе сайты – все элементы тренировочной выборки, связанные с магнием, несайты – случайная выборка без повторений из элементов, не имеющих связи с ионами магния, размер равен количеству сайтов.  
Общий размер тренировочной выборки уменьшается.

Оба этих подхода имеют свои достоинства и недостатки. Так, в первом случае существенно увеличивается вычислительная сложность алгоритма, в то время как во

втором случае с отсеиванием части несайтов, теряется некоторое количество информации о выборке.

Однако это не помешало второму подходу на кросс валидации показать результат, сравнимый с первым подходом. И поскольку первый способ требует существенно больше времени и ресурсов для работы RandomForest, в дальнейшей работе будет использоваться выборка сайтов, в которую случайным образом будет отобрано такое же число несайтов.

Для сравнения на рисунке (Рис. 4.2) ниже приведены аналогичные матрицы неточностей, построенные для последнего сбалансированного тренировочного разбиения выборки нуклеотидов с ионами магния, рассматриваемые на расстоянии до 5 Å от элементов выборки.

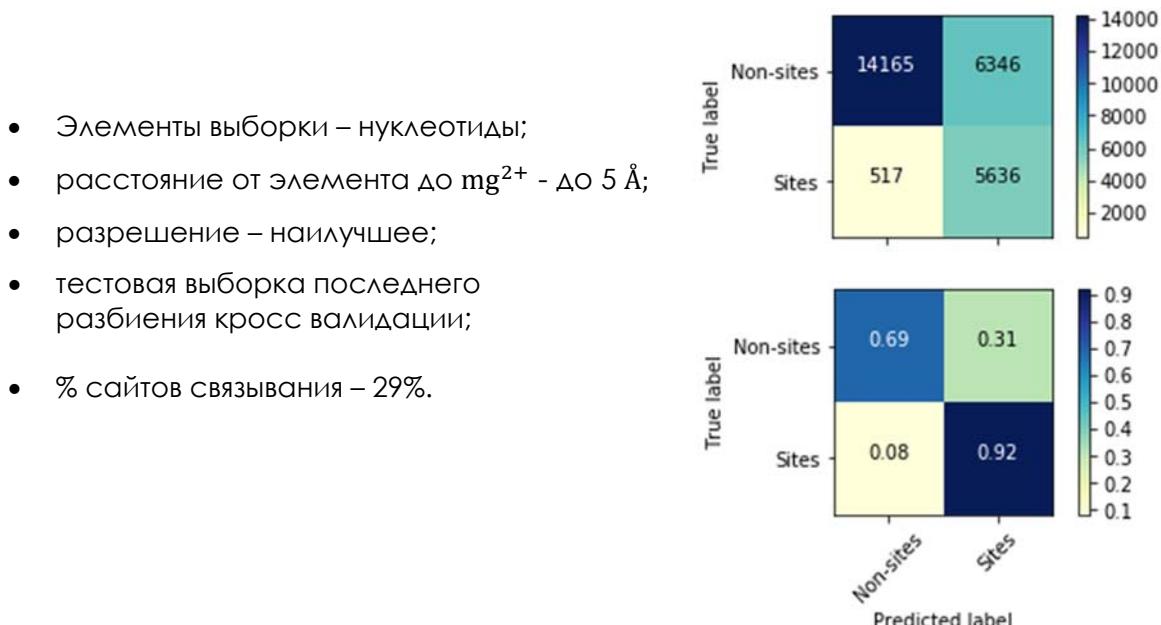


Рис. 4.2

Теперь необходимо включить балансировку в процесс кросс валидации. Это можно сделать двумя способами:

1. Сбалансиовать исходную выборку и далее делить ее на тренировочную и тестовую так, чтобы в этих выборках сохранялось соотношение количества сайтов к количеству не сайтов после балансировки, то есть данной соотношение было равно 1:1.
2. Делить исходную выборку на тренировочную и тестовую так, чтобы соотношение количества сайтов к количеству не сайтов было таким же, как и в исходной выборке. А далее балансировать тренировочную выборку.

Однако, в первом случае результаты кросс валидации могут оказаться смещанными и не отражать реального положения дел. Потому что каждый раз мы делим на тестовую и тренировочную выборку один и тот же сбалансированный набор, который к тому же существенно меньше нашей исходной выборки. Поэтому тестовые и тренировочные выборки будут иметь мало отличий друг от друга на разных итерациях в отличие от второго случая.

Здесь, на тренировочную и тестовую выборки делится исходная выборка, которая по размеру сильно превосходит выборку в первом случае. И далее происходит балансировка только тренировочной выборки, а тестовая сохраняет исходную долю сайтов. Таким образом, результаты кросс валидации получаются более объективными.

Таким образом, новая кросс валидация происходит по следующему алгоритму:

- в предоставленных наборах около 5% строк содержат пропущенные значения, на данном этапе эти строки будут опущены;
- каждый набор делится на тренировочную и тестовую выборки с исходной долей сайтов в выборах, при чем тестовая составляет около 30% от исходной выборки;
- тренировочная выборка балансируется так, чтобы число сайтов было равно исходному числу сайтов;
- для каждой пары выборок алгоритм RandomForest с параметрами, установленными по умолчанию, тренируется на первой выборке и делает предсказания на второй;
- для оценки качества предсказания вычисляется его точность: отношение числа элементов с правильно предсказанным целевым признаком к общему числу элементов;
- процедуры деления набора на выборки, тренировки алгоритма, предсказания и вычисления точности повторяются по 15 раз для каждого набора;
- таким образом, чтобы получить общую оценку работы алгоритма на наборе данных, точности, полученные на соответствующих 15 разбиениях, усредняются.

Ниже, в Таблица 4.3-4.4 приведены данные новой кросс валидации. Сравнивая эти результаты с предыдущими, можно отметить, что точность по всем выборкам снизилась, и зависимость, описанная для прошлой кросс валидации здесь не наблюдается.

**Таблица 4.3 Макс. число ионов**

Элементы	Радиус	%сайтов	Точность
Нуклеотиды	3 Å	8 %	0.71
Нуклеотиды	5 Å	29 %	0.61
Нуклеотиды	7 Å	42 %	0.55
Нуклеотиды	3-7 Å	41 %	0.56
Фрагменты	3 Å	3 %	0.74
Фрагменты	5 Å	14 %	0.69
Фрагменты	7 Å	27 %	0.63
Фрагменты	3-7 Å	27 %	0.64
Атомы	3 Å	0.9 %	0.8
Атомы	5 Å	7 %	0.78
Атомы	7 Å	19 %	0.73
Атомы	3-7 Å	18 %	0.73

**Таблица 4.4 Лучше разрешение**

Элементы	Радиус	%сайтов	Точность
Нуклеотиды	3 Å	6 %	0.69
Нуклеотиды	5 Å	24 %	0.6
Нуклеотиды	7 Å	36 %	0.55
Нуклеотиды	3-7 Å	36 %	0.56
Фрагменты	3 Å	2 %	0.77
Фрагменты	5 Å	11 %	0.68
Фрагменты	7 Å	22 %	0.63
Фрагменты	3-7 Å	21 %	0.63
Атомы	3 Å	0.7 %	0.83
Атомы	5 Å	5 %	0.77
Атомы	7 Å	15 %	0.72
Атомы	3-7 Å	15 %	0.72

При сравнении двух таблиц можно заметить несущественные отличия в точности предсказаний, поэтому далее лучше продолжать вести работу с данными наилучшего разрешения, так как в данных с худшим разрешением по определению будет находиться большее число шумов и неверно распознанных связей между атомами и ионами.

В выборках, в которых связи с ионами магния рассматривались в окрестности 3 Å, очень низкий процент содержания таких связей, а так как нам необходимо балансировать выборку, то тренировочная выборка в итоге получится очень маленькой, совершенно негодной для тренировки алгоритма RandomForest. Также можно заметить, что проценты содержания сайтов в выборках с ионами магния в окрестности 3-7 Å и 7 Å практически не отличаются, поэтому нет смысла отдельно рассматривать окрестность 3-7 Å.

Таким образом, выборку, с которой имеет смысл продолжать работу стоит искать среди выборок с наилучшим разрешением, в которых связи с ионами магния рассматриваются в радиусе 5, 7 Å.

Однако несмотря на то, что зависимость между % сайтов и точностью предсказания, присутствующая в результатах прошлой кросс валидации, в данном случае не наблюдается, корреляция между этими столбцами все равно присутствует. Так происходит потому, что средняя точность предсказаний вычисляется на несбалансированных тестовых выборках. Однако, эта метрика не подходит для выборок с большим дисбалансом классов, и плохо отражает действительность происходящего.

Поэтому, чтобы выбрать набор данных, с которым в последствии будет вестись работа, нужно смотреть на значения других метрик. Так, для оценки качества предсказаний на тестовой выборке мы будем опираться на графики гос-кривых, precision-recall, матрицы неточностей и распределения предсказаний.

Гос-кривая. Это график, показывающий соотношение между чувствительностью и специфичностью предсказаний в зависимости от порога для вероятности предсказаний, начиная с которого признается наличие связи нуклеотида/основания/атома с ионом магния.

При этом чувствительностью называется отношение числа правильно предсказанных сайтов связывания к числу реальных сайтов связывания, а специфичностью – отношение предсказанного числа не сайтов связывания к общему числу не сайтов.

Precision-recall – кривая. Это график, показывающий соотношение между точностью (precision) и полнотой (recall) предсказаний в зависимости от порога для вероятности предсказаний, начиная с которого признается наличие связи нуклеотида/основания/атома с ионом магния. Здесь точность (precision) совпадает с чувствительностью, а полнота – доля правильно предсказанных сайтов относительно общего числа предсказанных сайтов.

Таким образом, чтобы сравнить оставшиеся 6 выборок будет проведена 1 итерация кросс валидации, и для предсказанных вероятностей на тестовой выборке будут вычислены и построены 3 графика: гос-кривая, precision-recall – кривая и распределения, и 2 матрицы неточностей – исходная и нормированная (Рис. 4.4-4.5).

Ниже для сравнения приведена гистограмма точностей (accuracy), полученных на данной итерации кросс валидации (Рис. 4.3).

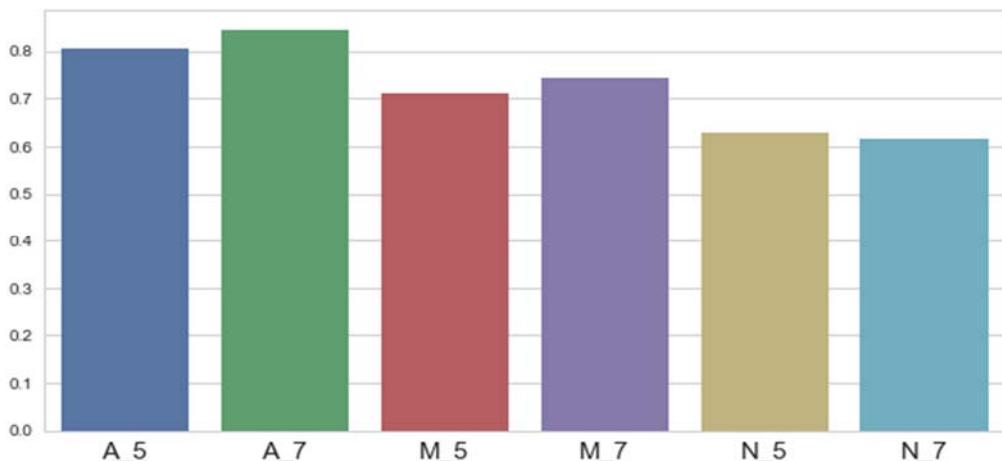


Рис. 4.3

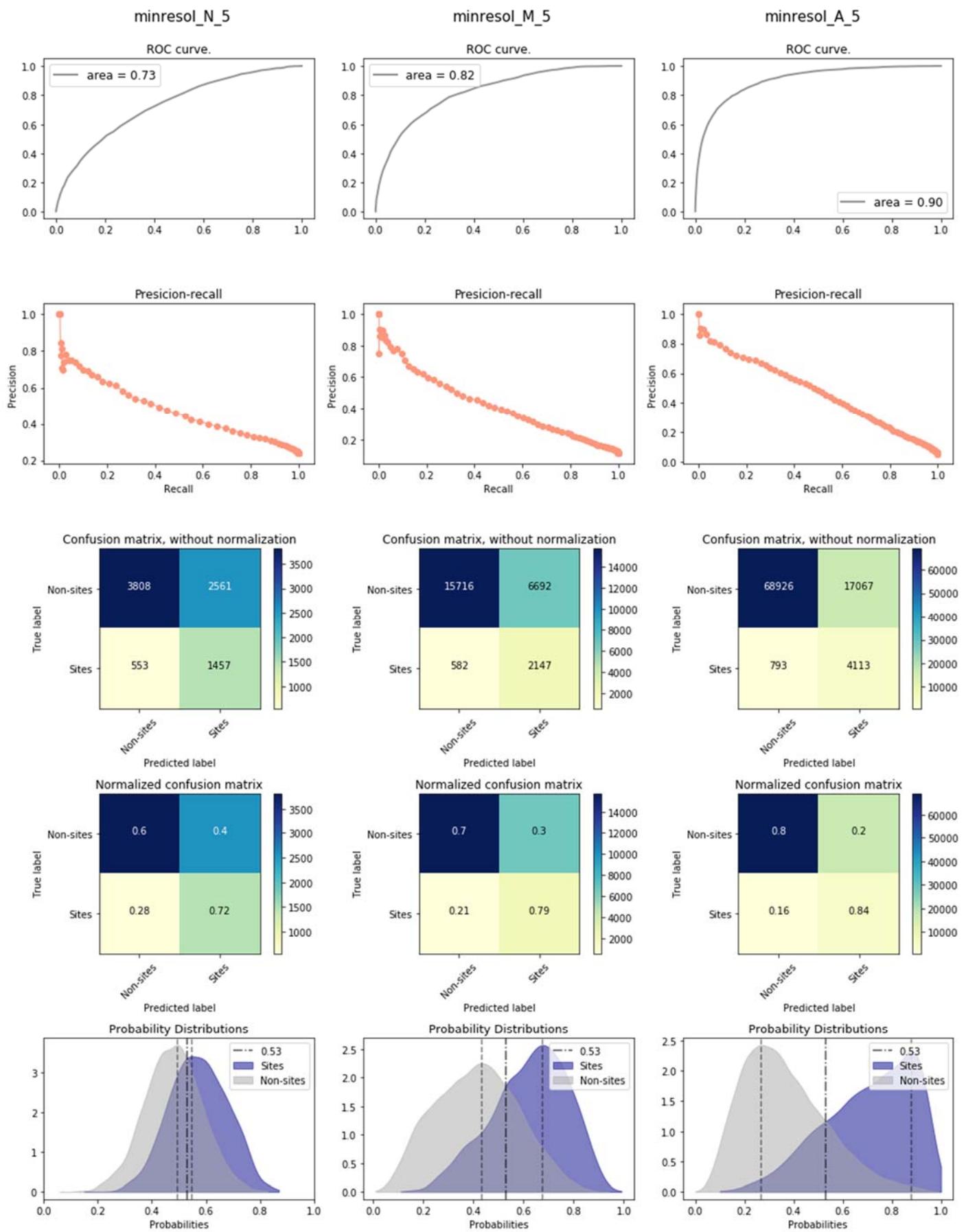


Рис. 4.4

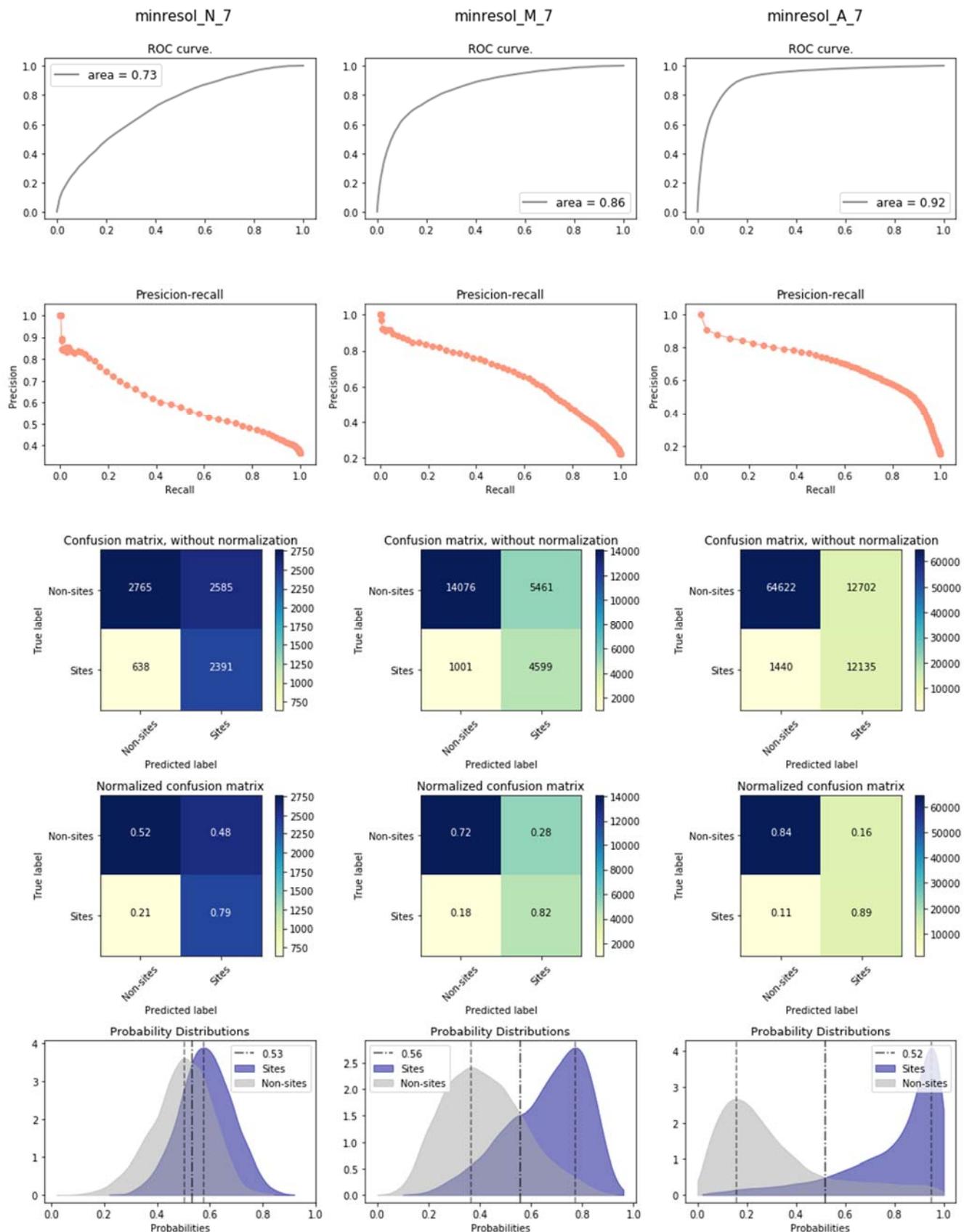


Рис. 4.5

Итак, судя по этим данным лучший результат показывает выборка minresol\_A\_7. С ней и будет продолжена работа в дальнейшем.

## 4.2 Препроцессинг/ Обработка данных

### 4.2.1 Пропущенные значения

На этапе предварительного анализа было обнаружено, что предоставленные наборы данных содержат около 5% строк с пропущенными значениями. Для первоначальной апробации модели это число можно было считать несущественным, поэтому было принято решение просто опустить эти строки. Однако, пришло время выяснить природу возникновения пропусков в данных: являются ли они дефектами в ходе проведения лабораторного исследования или же они являются отражением особенностей структур в данных?

Чтобы это выяснить, посмотрим на расположение пропусков в цепях (Рис. 4.6) и на распределение пропусков в признаках (Рис. 4.7 Рис. 4.6). Для наглядного представления расположения пропусков из рабочей выборки (`minresol_A_7`) было отобрано случайным образом 10 цепочек разной длины. Серыми линиями нарисованы сами цепочки, а синими точками отмечены пропуски в этих цепях.

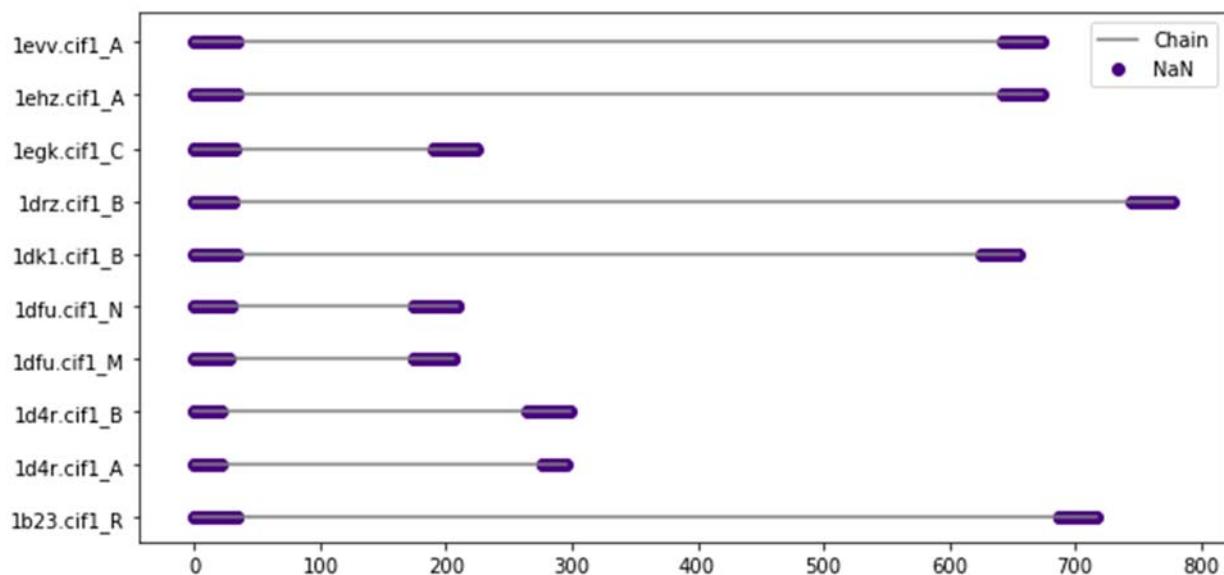


Рис. 4.6

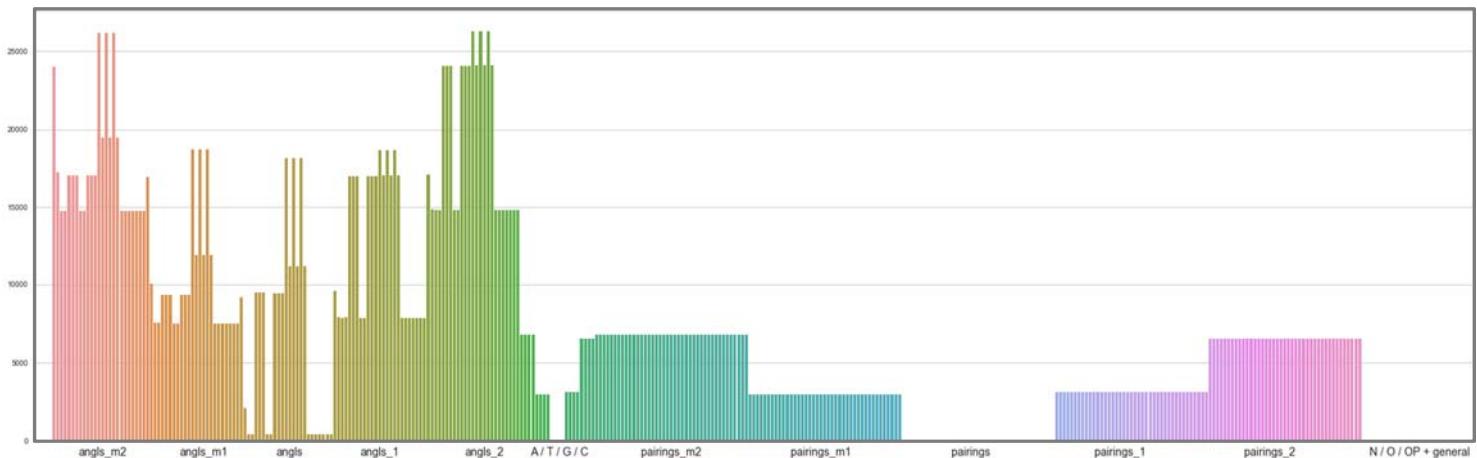


Рис. 4.7

По графику на Рис. 4.6 видно, что пропуски встречаются только на концах строк. Этот факт имеет вполне разумное и логичное объяснение, которое подтверждает и гистограмма распределения числа пропусков в признаках:

- В числе признаков, приведенных для каждого элемента, есть признаки, описывающие по 2 соседних элемента вверх и вниз по цепочке. Эти признаки: углы, спаривания и информация о нуклеотиде, из которого взят атом. Очевидно, что крайних нуклеотидов/атомов в цепочке нет соседних элементов с одной из сторон, а значит, для них невозможно описать указанные выше признаки.
- Помимо пропущенных значений в признаках, описывающих соседние нуклеотиды, пропуски имеются и в некоторых углах, приведенных для непосредственно рассматриваемого атома. Однако, причина этих пропусков также скрыта в особенностях расположения атомов в цепочке. Ведь значение угла с вершиной в данном атоме зависит от расположения соседних атомов. Таким образом, для крайних атомов мы тоже не можем вычислить значения признаков – углов, приведенных для этих элементов, ввиду отсутствия соседних элементов.

Итак, было выяснено, что пропуски в данных не случайны, а это значит, что игнорировать строки с пропущенными значениями, как это происходило раньше, – неверно. Необходимо сохранить информацию об этих элементах, учитывая знание об их расположении. Для этого нужно заполнить ячейки с отсутствующими значениями. Существуют различные стратегии, ниже приведены основные виды заполнений пропусков:

- нулями;
- соседними числами;
- средним/медианой по данному признаку;
- случайными числами;
- специальными числами, выходящими из области допустимых значений данного признака, либо в случае категориального признака – новой категорией.

Чтобы понять, какой способ заполнения является наиболее правильным в данном случае, нужно понять, что означает каждый из признаков, в котором встречаются пропущенные значения.

Мы уже определили, что пропуски встречаются в трех группах признаков: информации о нуклеотидах, спариваниях, углах. При этом пропуски первых двух групп встречаются только в признаках, приведенных для соседних нуклеотидов, а пропуски в углах могут быть для всех признаков.

1. Информация о нуклеотидах представлена в виде 4 бинарных признаков: seqA, seqC, seqU, seqG, каждый из которых может принимать значения 0 или 1, в зависимости от того, является ли рассматриваемый атом частью этого нуклеотида или нет. Таким образом, отсутствующий нуклеотид равносителен ситуации, когда все 4 признака имеют значение 0 одновременно.
2. Спаривания также представлены в виде 41 количественного признака: (*S<sub>I</sub>, S<sub>II</sub>, S<sub>III</sub>, ..., S<sub>XXIX</sub>, cWH, cWS, cWW, cHS, cHH, cSS, tWH, tWS, tWW, tHS, tHH, tSS*), каждый из которых может принимать целые неотрицательные значения, отражающие, какое количество спариваний данного типа имеет нуклеотид данного элемента выборки. Аналогично предыдущей группе признаков, отсутствующий соседний атом равносителен ситуации, когда все эти признаки имеют значение 0 (0 связей данного типа).
3. Углы представлены в виде вещественных признаков, показывающих непосредственно значение угла с вершиной в данном или соседнем атоме.

Заполнение этих признаков нулями для отсутствующих атомов также не противоречит сути признаков.

Таким образом, получается, что в случае с нашими данными оптимальным и допустимым способом работы с пропущенными значениями, является заполнение соответствующих ячеек нулями.

## 4.2.2 Анализ признаков

После того, как все пропуски в наших данных ликвидированы, необходимо оценить скоррелированность признаков, а также их заполненность.

Для наглядности ниже на Рис. 4.8 приведена тепловая карта рассматриваемого датасета. Здесь строки – атомы, столбцы – признаки, значения нормируются внутри каждого признака. На рисунке видно, что в наших данных большое количество разреженных столбцов, то есть таких признаков, которые имеют одно и то же значение почти для всех атомов. В основном это спаривания.

Очевидно, что такие признаки имеют малую информативность для алгоритма распознавания сайтов связывания, но существенно увеличивают его вычислительную сложность. Поэтому, стоит исключить такие признаки из рассмотрения: для каждого столбца вычисляется его дисперсия, затем столбцы с дисперсией ниже некоторого порога удаляются из выборки. Для данной выборки был установлен порог, равным 0.004. Этот порог выбирался эмпирическим путем так, чтобы результат визуально соответствовал действительности.

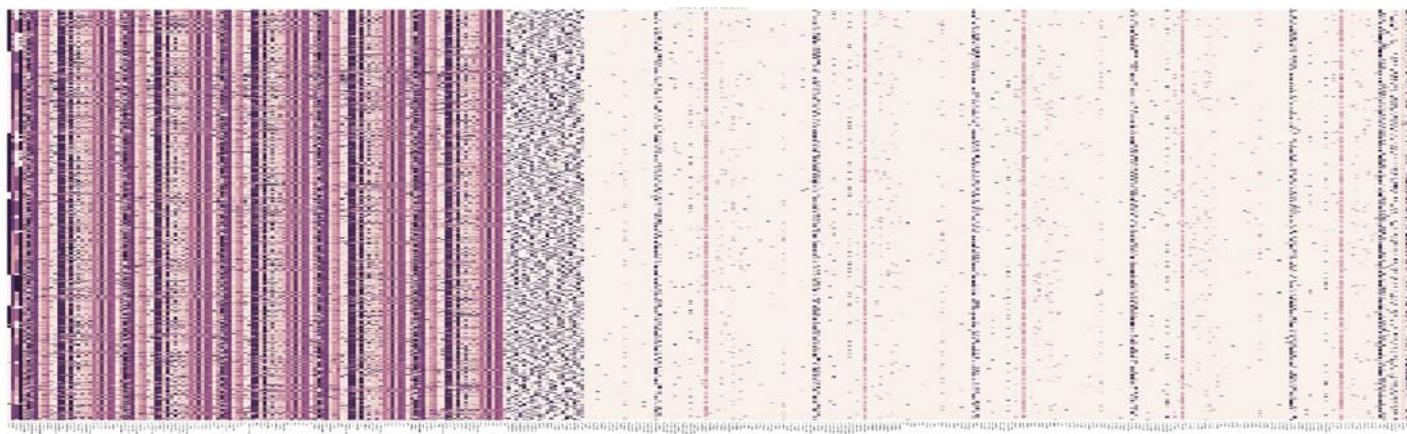


Рис. 4.8

Далее оцениваем скоррелированность оставшихся признаков: вычисляем матрицу корреляций и, чтобы можно было проанализировать полученные данные, построим тепловую карту этой матрицы (Рис. 4.9). Вычисление матрицы корреляций и построение тепловой карты был сделано методами языка R, который также позволяет строить дендрограммы на основе полученных значений и отсортировать матрицу корреляций в соответствии с этими дендрограммами.

Так, по левой и верхней осям тепловой карты изображены дендрограммы, а по нижней и правой оси названия соответствующих признаков. Также на этой тепловой карте отмечены области наиболее скоррелированных групп признаков, которые будут рассматриваться в дальнейшем более подробно.

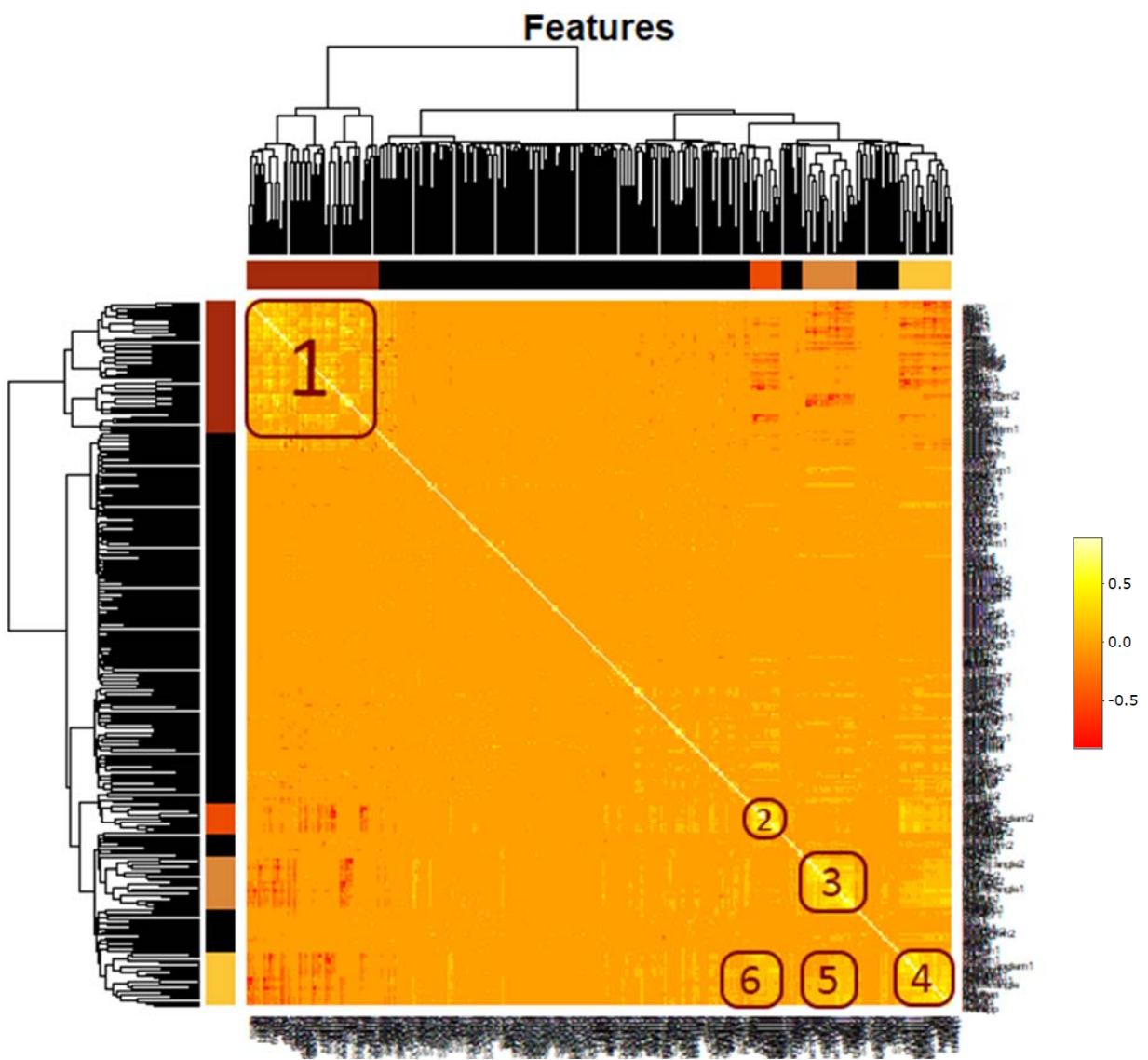


Рис. 4.9

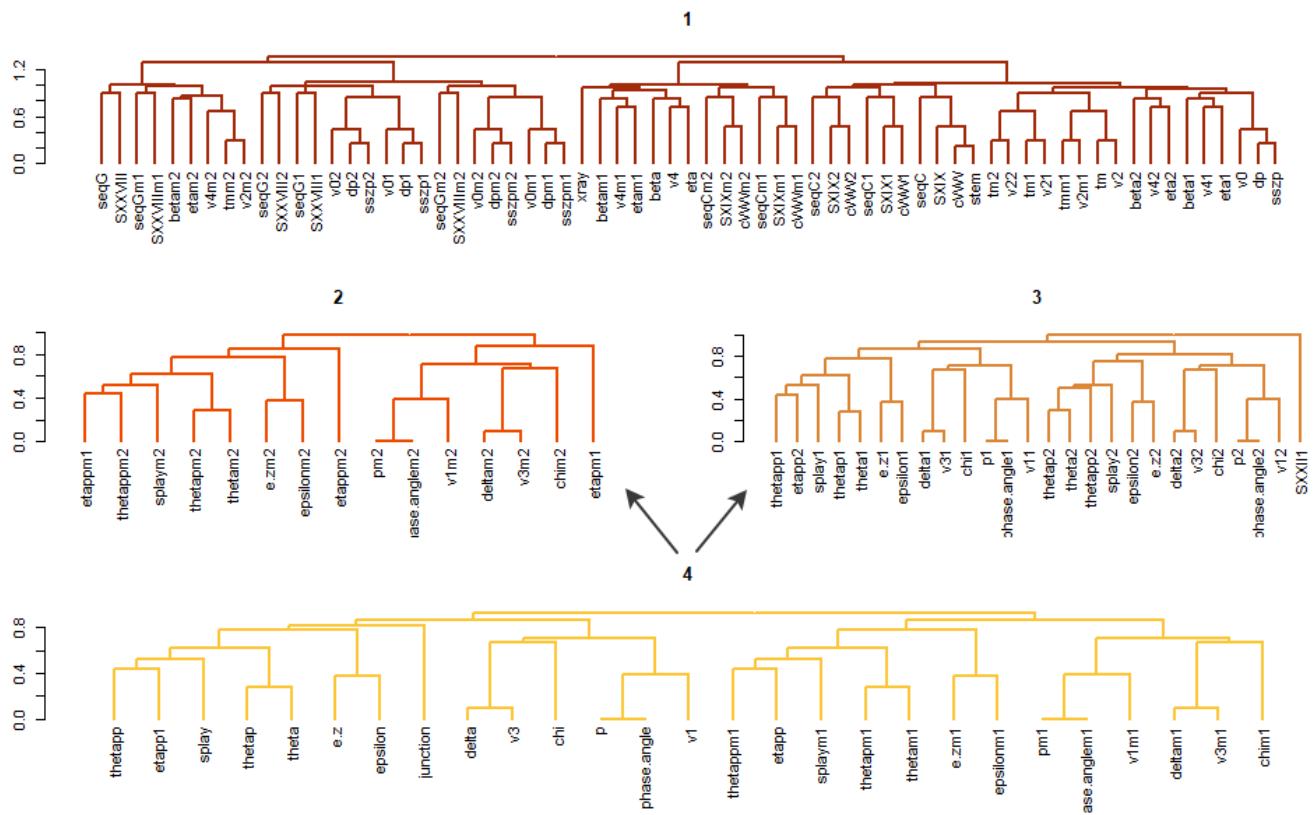


Рис. 4.10

На Рис. 4.10 отображены дендрограммы признаков наиболее скоррелированных групп. Стрелочками показаны корреляции между группами признаков. Сравнивая дендрограммы с тепловой картой нетрудно увидеть, что корреляции подвержены только углы. Поэтому для дальнейшего выявления групп корреляций были произведены аналогичные вычисления, но только для углов атомов. Тепловая карта и дедрограммы наиболее скоррелированных областей приведены ниже (Рис. 4.11-4.12).

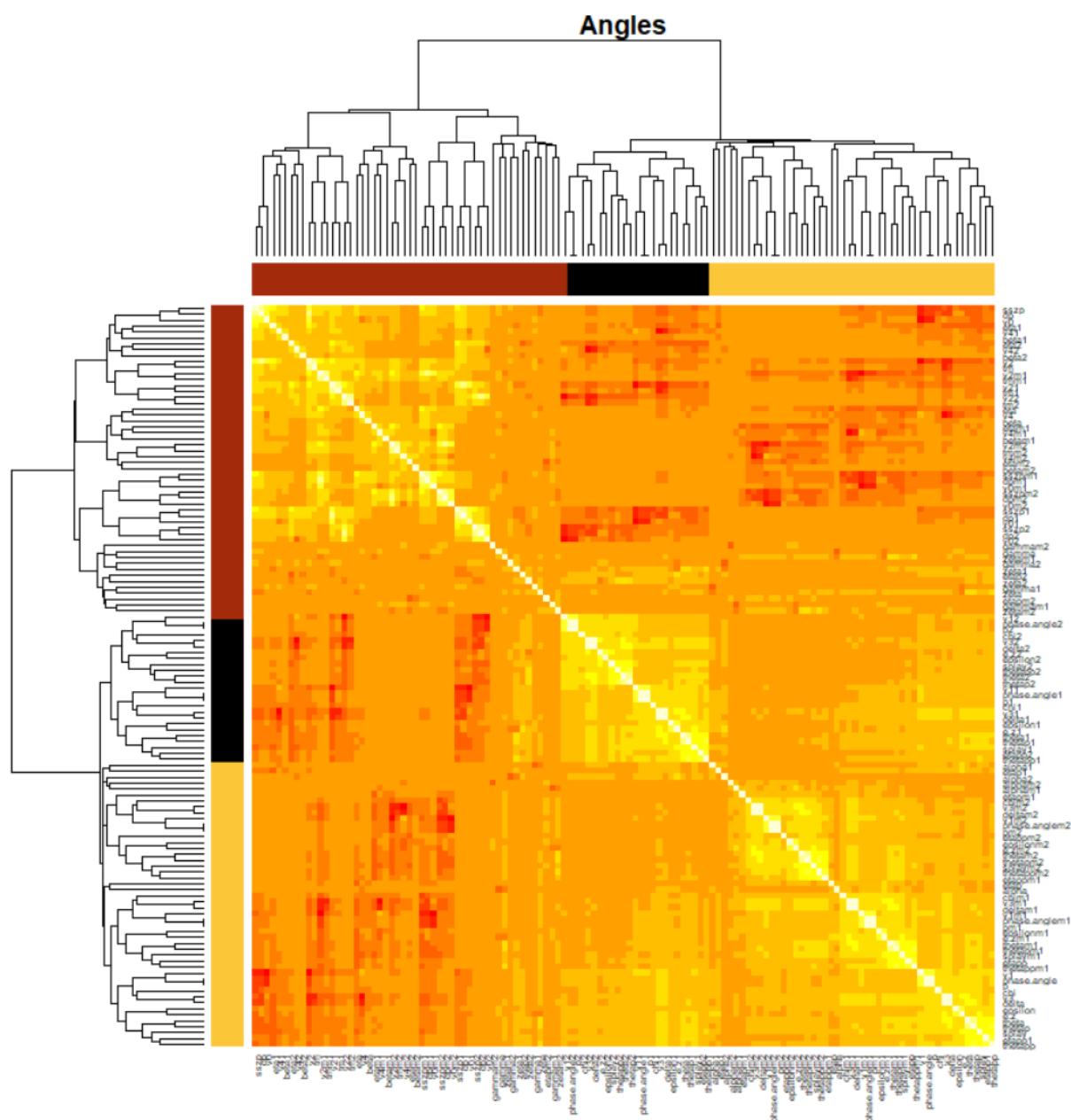


Рис. 4.11

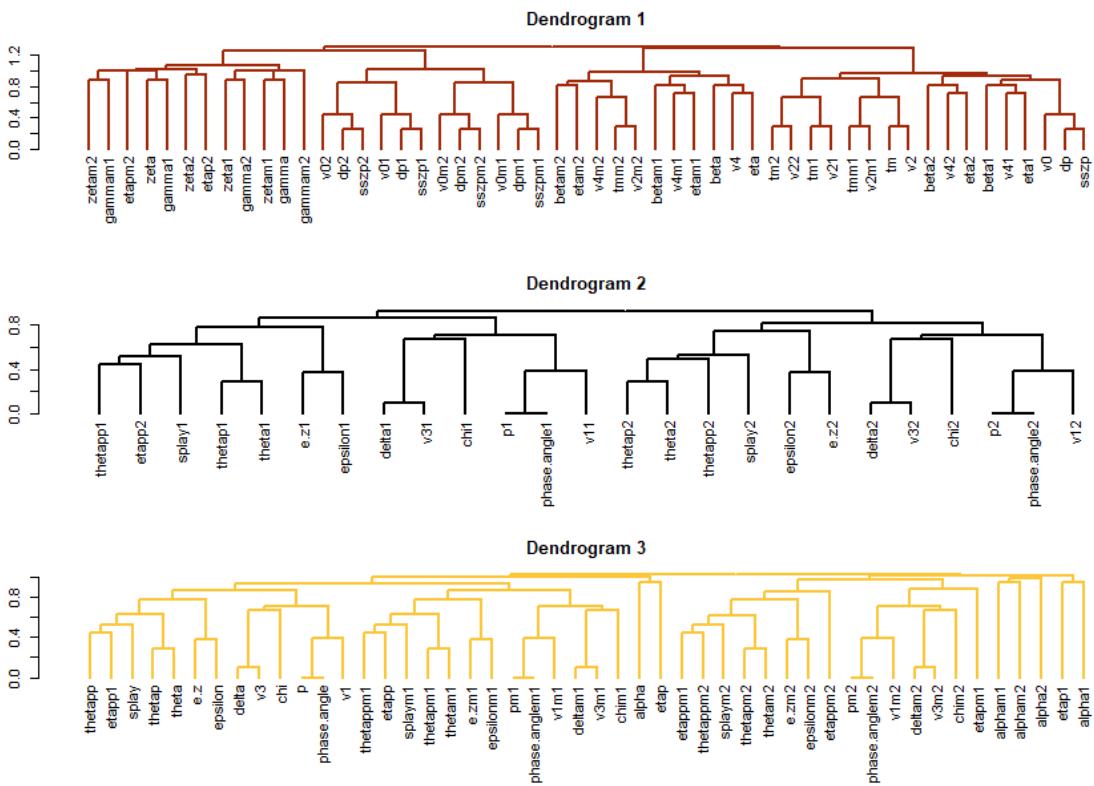


Рис. 4.12

Анализируя полученные дендрограммы и сопоставляя их с тепловой картой корреляций, были выведены следующие группы наиболее скоррелированных признаков:

- thetaapp, etapp1, splay, thetap, theta;
- e.z, epsilon;
- v0, dp, sszp;
- v1, p, phase.angle;
- v2, tm;
- v3, delta.

Таким образом, для дальнейшей работы из каждой группы углов будет выбрано по одному признаку, а остальные углы будут удалены из выборки. Аналогичные действия будут произведены с углами, описанными для двух соседних нуклеотидов вверх и вниз по цепочке.

### 4.2.3 Анализ структур в выборке

По результатам общения с кристаллографами было выяснено, что лабораторные исследования с длинными цепочками проводить существенно сложнее, чем с короткими, поэтому в таких структурах достаточно высокий процент ошибок. Это означает, что в ходе нашей работы лучше исключить из рассмотрения такие структуры, поскольку они содержат большое количество шумов, что неблагоприятным образом скажется на работе алгоритма.

## 4.3 Подбор параметров алгоритма

---

Теперь, когда данные приведены к нужному формату и отобраны необходимые признаки, можно переходить к настройке параметров алгоритма. Основные параметры, которые будут подбираться так, чтобы позволять алгоритму обучаться наилучшим образом на наших данных:

- доля признаков, которые отбираются случайным образом на каждом шаге построения дерева;
- глубина деревьев;
- минимальное число элементов в листе.

Выбор наилучших параметров будет осуществляться полным перебором из заданных диапазонов значений, при этом лучшим набором параметров будет считаться тот, с которым RandomForest покажет самое высокое качество предсказания на кросс валидации.

Процесс кросс валидации в данном случае будет немного отличаться от того, который проводился для оценки потенциала предсказания предоставленных наборов данных.

Главные отличия:

- Деление на тренировочную и тестовую подвыборки. Так как в нашем случае набор данных представляет собой набор структур, описанных на уровне атомов и, в конечном счете, ионы магния влияют на конфигурацию структуры в целом, то деление на подвыборки лучше проводить так, чтобы не перемешивать атомы одной структуры между тренировочной и тестовой выборками. Это позволит исключить возможное влияние тренировочной выборки на тестовую и сделает процедуру кросс валидации более объективной.
- Метрика качества предсказания. В прошлый раз в качестве метрики предсказания сайтов связывания вычислялась точность (accuracy) предсказаний. Однако, как мы уже могли убедиться, данная метрика плохо работает на несбалансированной выборке. Например, когда мы выбирали лучшую для предсказаний выборку, эта метрика была не способна отразить тот факт, что алгоритм не научился предсказывать класс, представленный малым числом объектов. В таких случаях лучше опираться на другие метрики: точность (precision) и полноту (recall).

Здесь точность – это доля правильно предсказанных сайтов относительно общего числа реальных сайтов, полнота – доля правильно предсказанных сайтов относительно общего числа предсказанных сайтов. Однако, оценивать две метрики параллельно довольно затруднительно, поэтому мы будем рассматривать метрику  $f_1$  – гармоническое среднее между точностью и полнотой.

Таким образом, процесс кросс валидации для подбора параметров имеет следующий вид:

1. Загрузка и препроцессинг выборки minresol\_A\_7.
2. Формирование всех возможных наборов параметров из заданных диапазонов значений.
3. Для каждого набора параметров:

- Выборка делится на тренировочную и тестовую подвыборки так, чтобы атомы одной цепочки оказались в одной подвыборке, при этом тестовая составляет около 30% от исходной выборки.
- Тренировочная выборка балансируется так, чтобы число не сайтов было равно исходному числу сайтов.
- Для каждой пары выборок алгоритм RandomForest с соответствующим набором параметров, тренируется на первой выборке и делает предсказания на обеих выборках.
- На предсказаниях тренировочной и тестовой подвыборок вычисляется f1-мера.
- Шаги a-d повторяются 5 раз.
- Усреднение значений f1-меры на тренировочной и тестовой выборках по всем разбиениям для данного набора параметров.

По результатам вышеописанного процесса был получен следующий график (Рис. 4.13). Здесь наборы параметров отсортированы по возрастанию среднего значения f1-меры на тренировочных подвыборках.

Также для более полного понимания происходящего на каждом шаге итераций сохранялись значения точности (precision), полноты (recall) и точности (accuracy), они также отображены на графике.

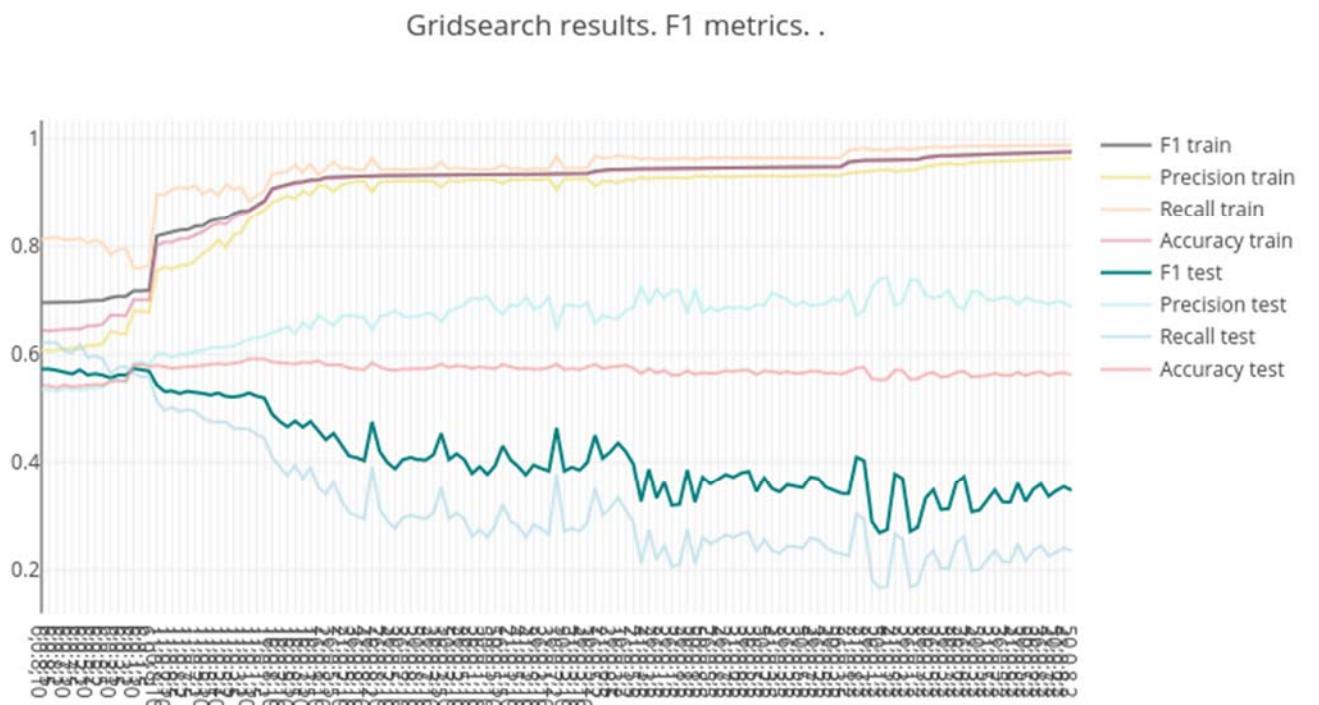


Рис. 4.13

Анализируя результаты кросс валидации гиперпараметров, можно отметить что при низких значениях `max_depth`, алгоритм RandomForest не успевает обучаться. При этом на тестовой выборке точность понемногу увеличивается, в то время как полнота с ростом глубины дерева начинается снижаться. Это значит, что деревья с глубиной 20-30 вершин уже достаточно обучились, чтобы относительно точно предсказывать некоторую часть сайтов связывания, но не являются переобученными. Также стоит отметить, что пики полноты коррелируют с провалами в точности и наоборот. И это объяснимо: по всей видимости в моменты пиков полноты алгоритм выделяет слишком

много сайтов. А в большее число сайтов, с равной вероятностью попадет и большее число как реальных сайтов (увеличение полноты), так и не сайтов (снижение точности). Аналогичная ситуация с моментами пиков точности. Эти пики приходятся на деревья, у которых в листах меньше 6 элементов. Такие деревья являются глубоко обученными, а в случае с RandomForest, где работа ведется с целым ансамблем таких деревьев, нет необходимости работать с ними, тем более, что данном случае, как мы видим, они мало полезны. Также, обычно, в RandomForest признак max\_features по умолчанию равен  $\sqrt{n}$ , где n - общее число признаков. Однако, мы имеем дело с большим числом разреженных признаков, поэтому этот параметр стоит выбрать достаточно большим.

Таким образом, наилучшим набором гиперпараметров будет считаться тот, при котором достигается некоторый оптимум в обучении деревьев и последующих предсказаниях с учетом особенностей нашей выборки.

Этот набор: max\_depth = 26, max\_features = 0.7, min\_samples\_leaf = 20.

## 4.4 Алгоритм построения модели

Таким образом был собран процесс обучения алгоритма RandomForest и обработки данных, при котором отбираются только самые значимые признаки, содержащие существенное количество информации. Также данные приводятся к виду, с которым позволяет работать данная реализация алгоритма RandomForest.

Ниже этот процесс расписан пошагово.

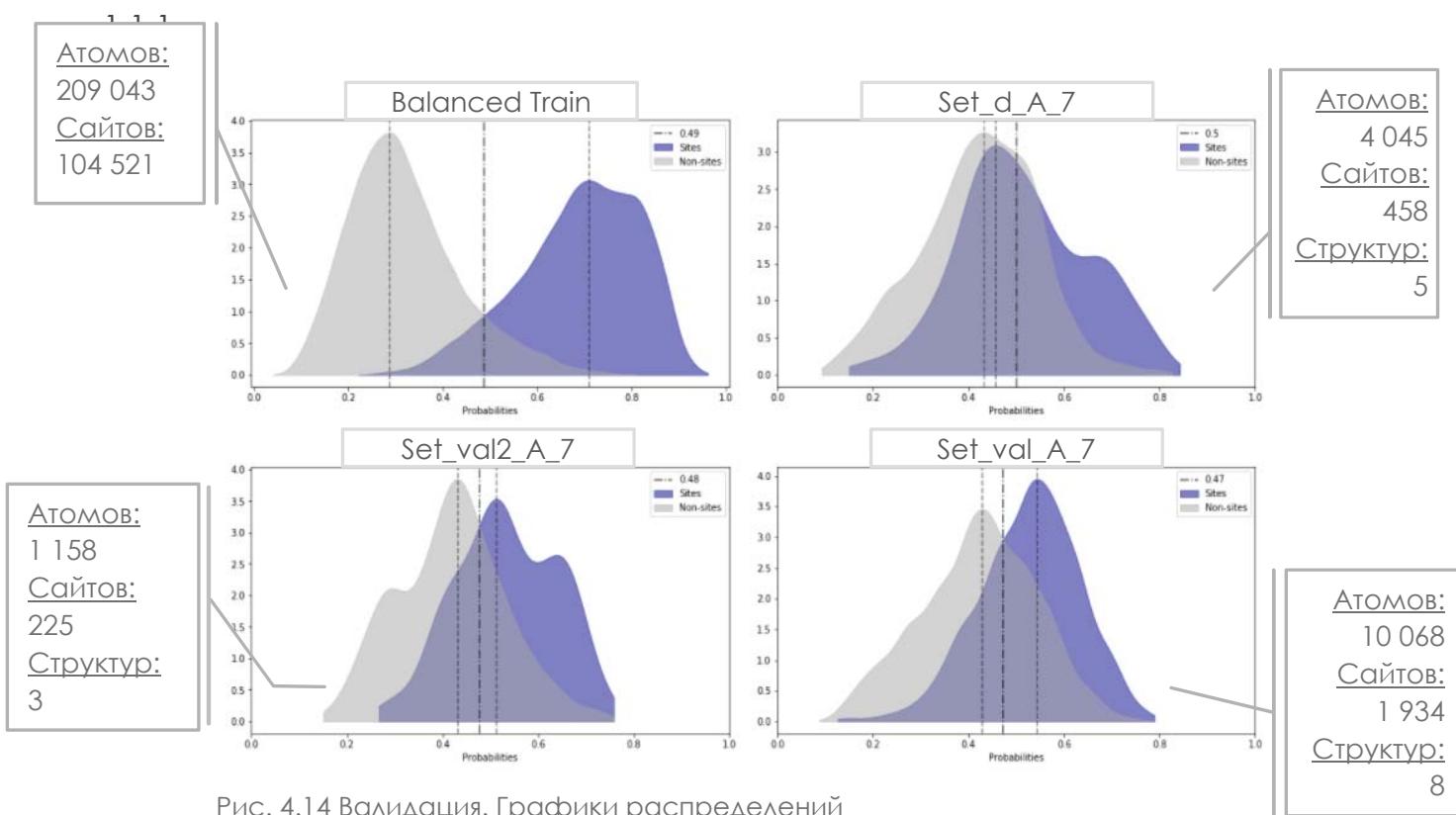
1. Препроцессинг данных (тренировочной и тестовой выборки).
  - a. Заполнение пропущенных значений нулями.
  - b. Удаление признаков с маленькой дисперсией.
  - c. Удаление скоррелированных признаков.
2. Тренировка модели.
  - a. Балансирование выборки: элементы – не сайты связывания отбираются случайным образом в количестве, равном числу сайтов связывания.
  - b. Тренировка RandomForest с параметрами:
    - Max\_depth = 26
    - Min\_samples\_leaf = 20
    - Max\_features = 0.7
3. Предсказание натренированной модели на тестовой выборке.

## 4.5 Валидация модели

Технический прогресс не стоит на месте: в лабораториях периодически меняется оборудование, растут темпы обработки информации, а вместе с тем растут и требования к объемам анализируемой информации, правда не всегда эти изменения положительно сказываются на качестве лабораторных исследований. Поэтому, чтобы оценить качество предлагаемого алгоритма имеет смысл смотреть на структуры, полученные в разные времена.

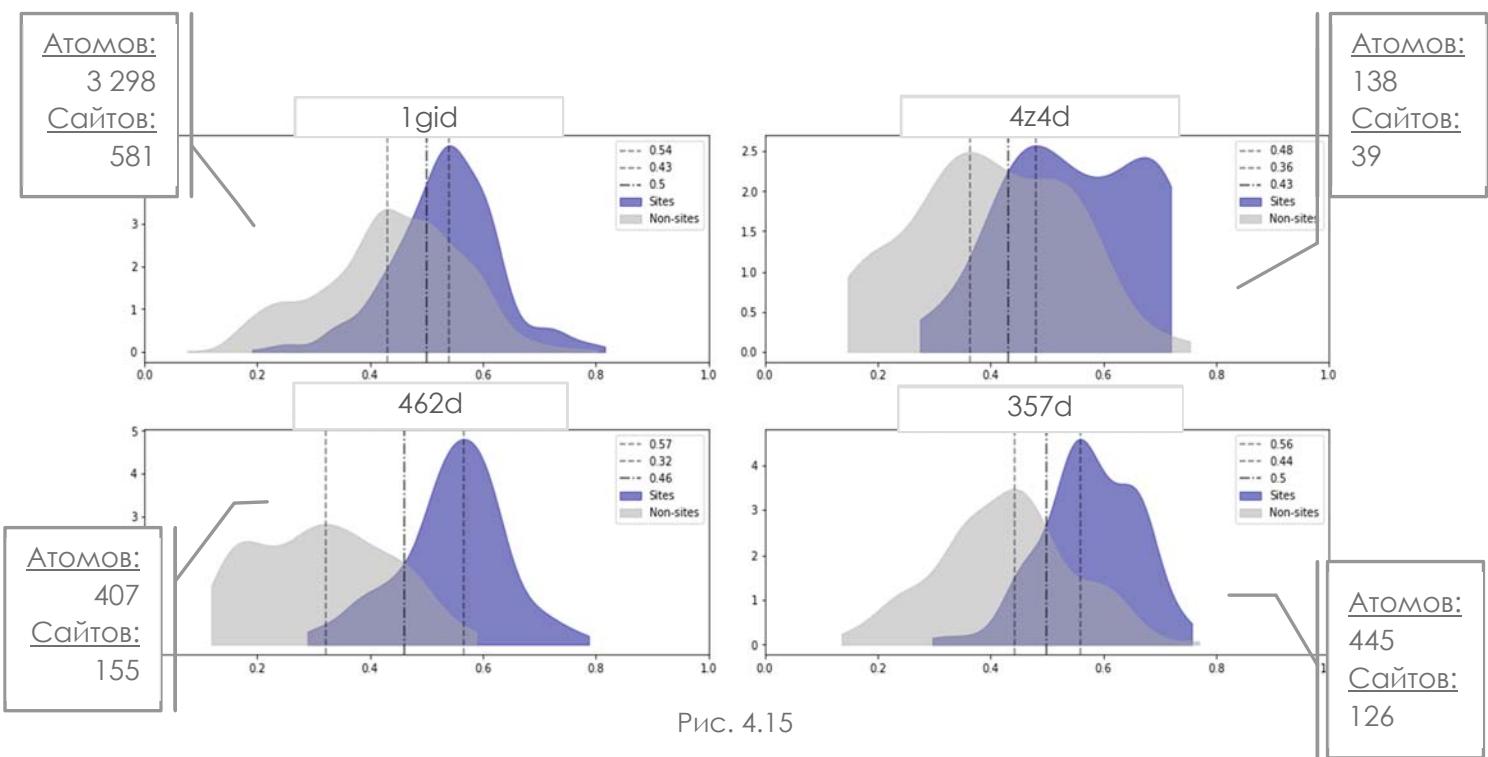
Наша исходная рабочая выборка, содержит 242 структуры, полученные в период от 2003 до 2017 года. Для валидации структур этого периода отберем случайным образом 5 структур из этой выборки, которые будут исключены из исходной рабочей выборки. Они образуют набор данных, который в этой работе будет фигурировать под названием *set\_d\_A\_7*. Также из базы данных *URS DataBase* [10] было отобрано 3 наиболее подходящих новейших структуры, опубликованных в базе данных позднее, чем была сформирована тренировочная выборка. Эти 3 структуры образуют набор *set\_val2\_A\_7*. В качестве набора структур *set\_val\_A\_7*, полученных раньше 2003 года, были взяты структуры, которые использовались для тренировки алгоритмов сервиса *Web-Feature*, так как для этой работы отбирались наименее зашумленные молекулы [7].

Итак, для валидации было сформировано 3 набора данных: *set\_d\_A\_7*, *set\_val\_A\_7*, *set\_val2\_A\_7*. Качество предсказаний будет оцениваться по графикам распределений. Также для контроля тренировки самой модели графики распределений были построены и для тренировочной сбалансированной выборки. Эти 4 графика распределений, а также небольшая статистика по наборам для валидации изображены на Рис. 4.14.



На графиках распределений для выборок структур от 2003 года заметно, что некоторая часть сайтов классифицируется с вероятностью около 0.7, что соответствует результату на тренировочной выборке. А в выборке структур до 2003 года пик приходится на 0.53. Это значит, что среди всех предсказаний в наших выборках есть структуры, для которых предложенный алгоритм сравнительно неплохо классифицирует сайты связывания. На Рис. 4.15 показаны некоторые из них.

В следующей части на основе этих предсказаний будут построены визуализации для некоторых цепочек из валидационных выборок, чтобы сравнить предложенный алгоритм с существующими онлайн-сервисами. И также, для сравнения, одна из цепочек с наиболее точно предсказанными сайтами связывания будет вычислена с помощью онлайн-сервисов.



## 5. Результаты

### 5.1 Оценка работы текущих сервисов

MetalionRNA – сервис, принцип работы которого разбирался в разделе 2 «Анализ существующих алгоритмов» был создан специально для предсказания связей ионов магния со структурами РНК. Создатели описывают довольно высокое качество предсказаний ( $\text{roc-auc} \approx 95\%$ ) и гарантируют высокую скорость обработки данных (для цепочки длиной порядка 76 нуклеотидов сервис вычислит расположение ионов магния менее, чем за 5 мин) [11].

Однако, этот сервис является однопоточным, поэтому, если им одновременно начинает пользоваться большое количество человек или загружать данные большого объема, то время обработки ваших данных может существенно увеличиться. Так, например, для оценки работы этого сервиса была загружена структура 1hr2, она была 10ая в очереди, однако, вычисления заняли около 4 месяцев. Позже нагрузка на сервер спала, и действительно, в течение часа можно было получить результат вычислений.

Изначально создававшийся для работы с цепочками ДНК онлайн-сервис WebFeature к моменту написания этой работы закрылся на неопределенное время. Поэтому, к сожалению, сравнить время его работы с MetalionRNA и качество предсказаний на определенных цепочках не удастся.

Для сравнения в начале посмотрим на структуры, описанные в самих статьях [3] [11]: 1hr2 (P4-p6 Domain Of Tetrahymena Thermophila Group I Intron), 1qa6 (58 Nt Sequence rRNA From Escherichia Coli) и 1hc8 (Bacillus Stearothermophilus 23s rRNA). К сожалению, в данных, предоставленных для этой работы структуры 1qa6 не нашлось, а структура 1hc8 не описана в статье о методе WebFeature. Таким образом, сравнение предсказаний будет вестись между двумя онлайн-сервисами: WebFeature и Metallon, и алгоритмом, предложенным в этой работе.

Вследствие того, что WebFeature на данный момент недоступен, то все приведенные ниже картинки для этого сервиса были взяты из статьи [3]. Все 3D визуализации были построены с помощью программы PyMol.

Стоит напомнить, что в онлайн-сервисах по концентрации предсказанных сайтов связывания вычисляются точные координаты предполагаемого местонахождения иона магния. Однако алгоритм, предложенный в данной работе, пока что может определить только предполагаемые сайты связывания ионов магния с атомами структуры, поэтому визуализации предсказаний этого алгоритма будут несколько отличаться от визуализаций онлайн-сервисов.

Итак, в визуализациях (Рис. 5.1-5.4), основанных на предсказаниях Metallon и WebFeature, желтым цветом показано расположение ионов магния, полученное экспериментальным путем, а красным цветом – предсказанное сервисом расположение ионов. В визуализациях алгоритма RandomForest экспериментально вычисленные расположения ионов магния показаны темно-зеленым цветом, а желтым цветом – показаны атомы, находящиеся в радиусе 7 Å от ионов, то есть целевой признак нашей выборки. Оранжевым цветом указаны атомы, которые RandomForest верно классифицировал, как близко расположенные к иону магния, светло-зеленым – атомы, которые алгоритм посчитал, как близко расположенные к иону магния, хотя они таковыми не являются.

Рис. 5.1 1hr2

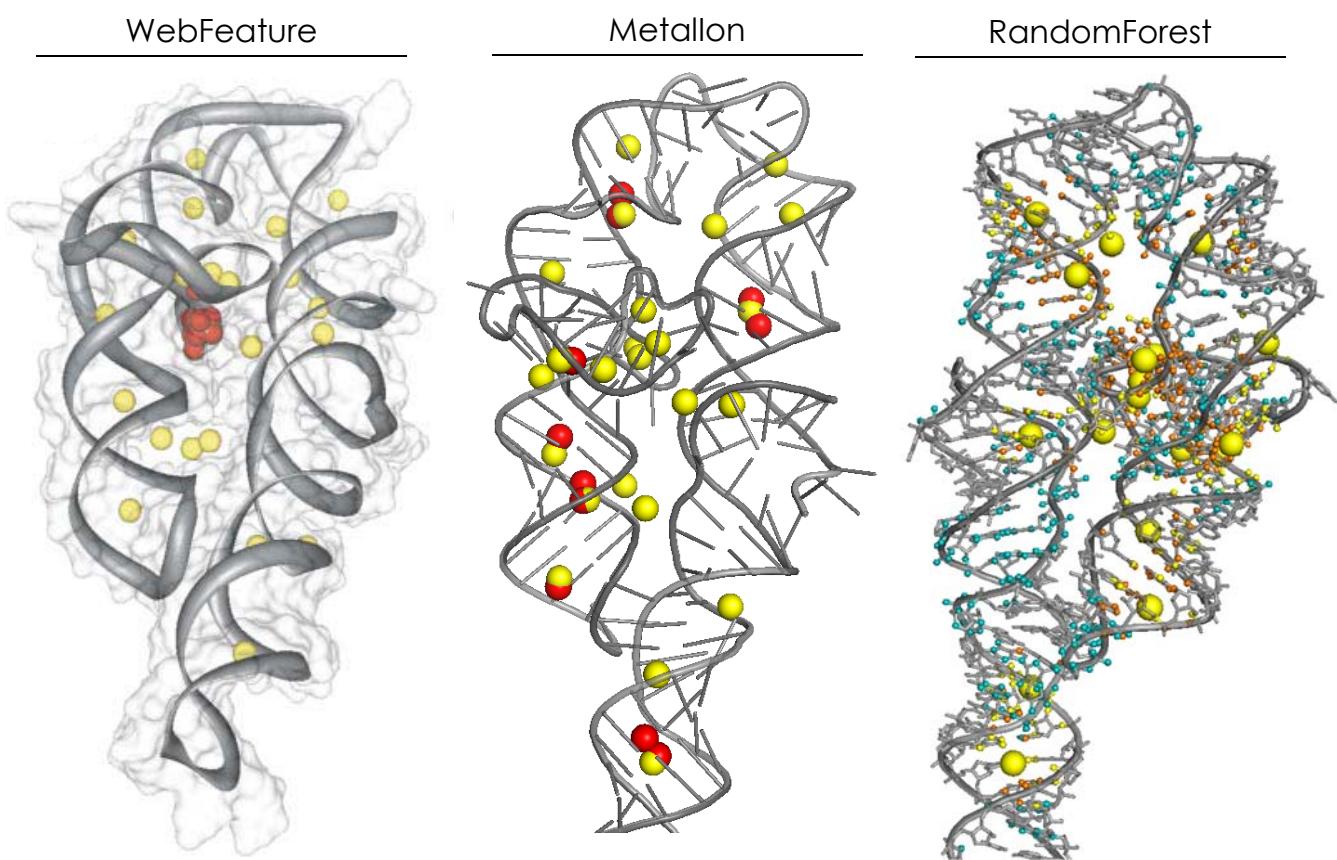


Рис. 5.2 1hc8

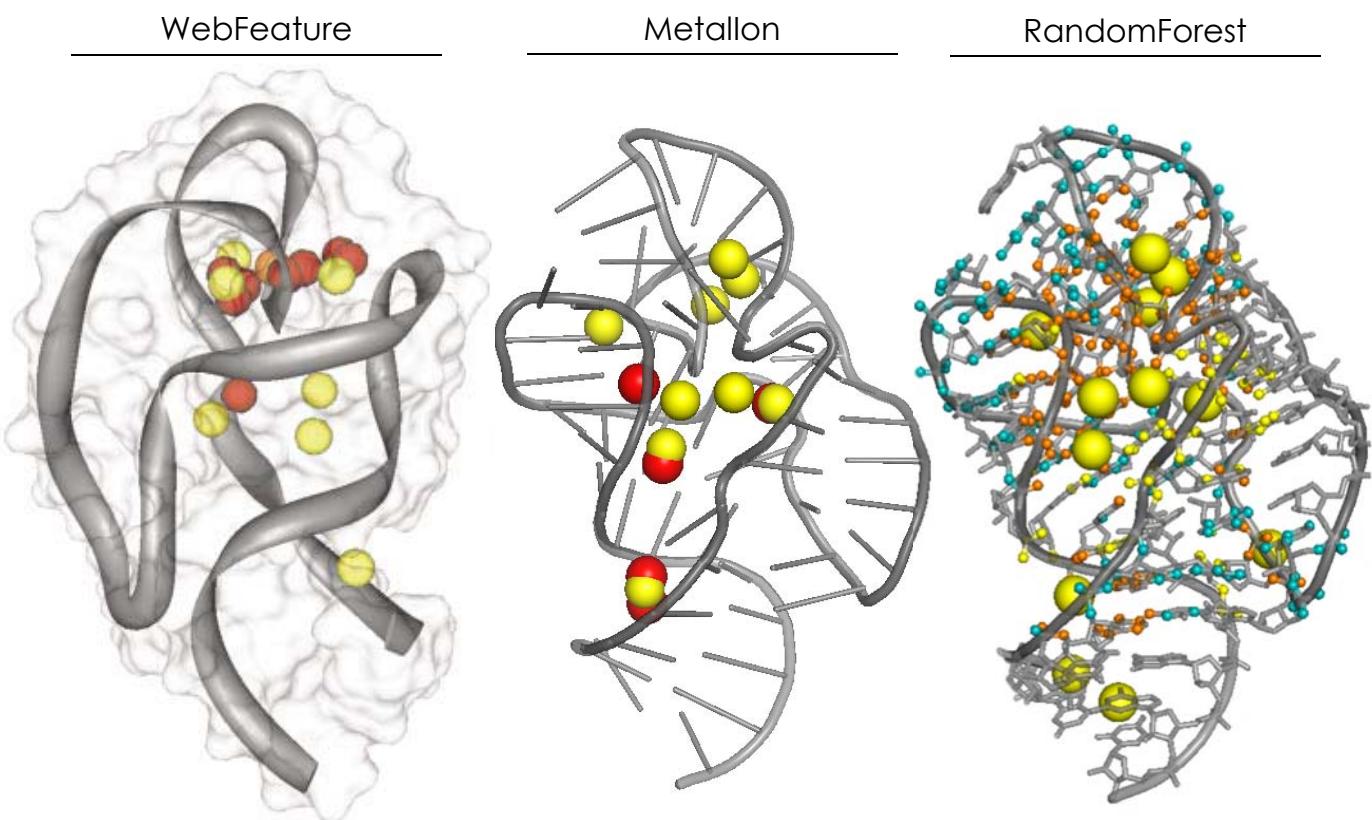


Рис. 5.3 1qa6

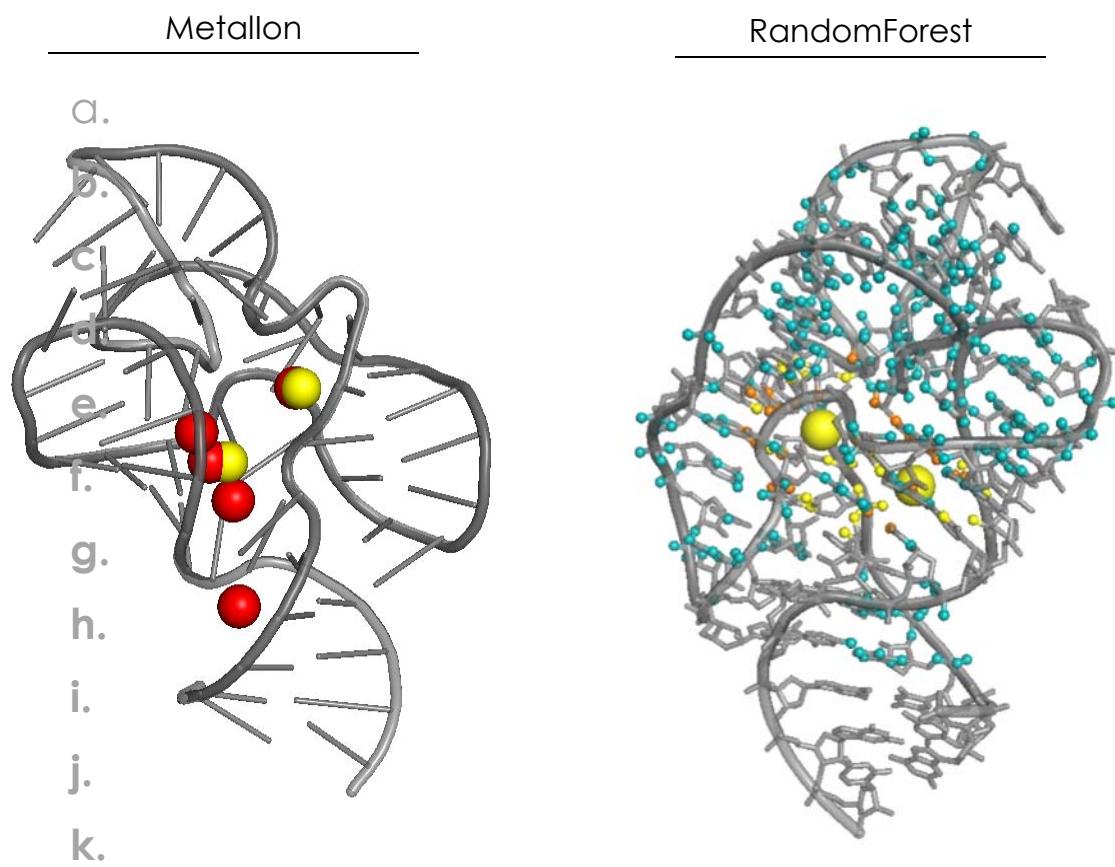
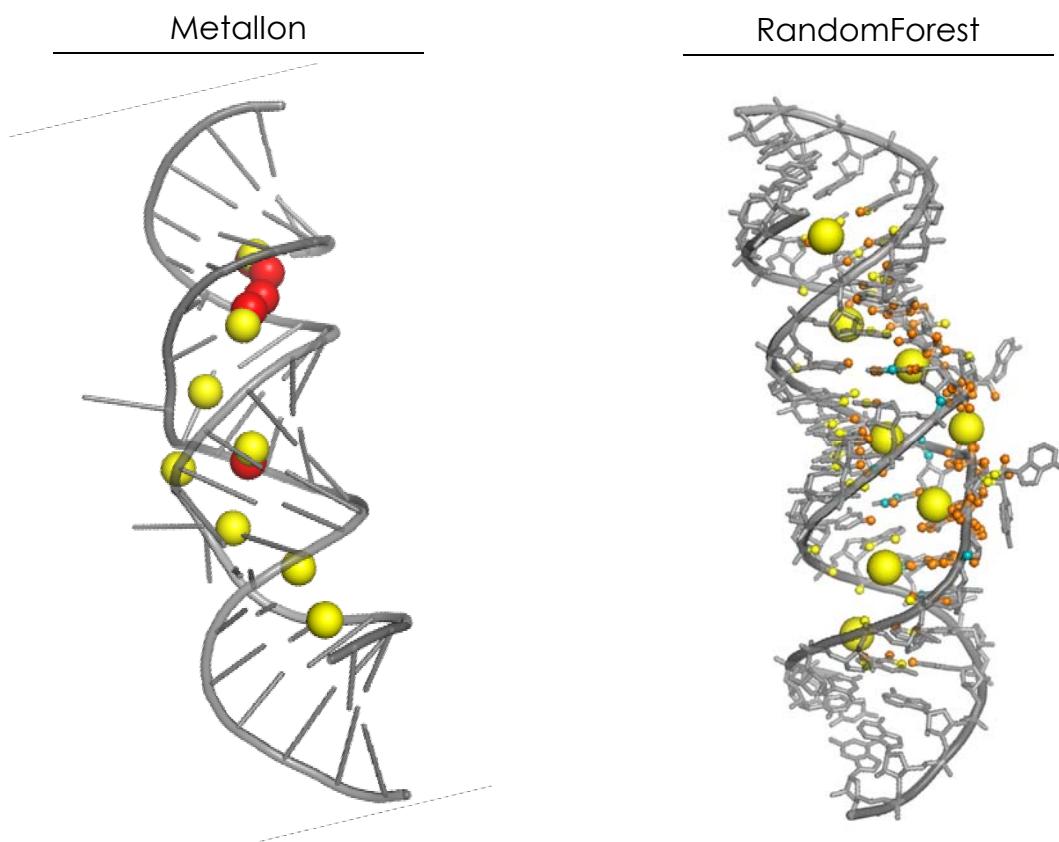


Рис. 5.4 462d



Сравнивая полученные визуализации, можно отметить, что WebFeature хуже справляется с задачей распознавания сайтов, чем Metallon. Зато работа RandomForest вполне сравнима с работой Metallon. Он, так же как и Metallon, немного избыточен, но области вокруг большинства ионов магния, а особенно вокруг скопления ионов магния, разработанный алгоритм распознает. Однако для демонстрационной структуры из статьи Metallon [11] - 1qa6, этот алгоритм почти не распознал сайтов связывания, но в то же время Metallon находится в аналогичной ситуации со структурой, с которой RandomForest прекрасно справляется (462d).

## 5.2 Апроксимация координат ионов магния

---

На данный момент модель разработана так, чтобы показывать атомы молекулы РНК, рядом с которыми (в радиусе 7 Å) находятся ионы магния. Однако, в случае, когда ионы атомов расположены довольно близко друг от друга, множества атомов, на которые они оказывают влияние, пересекаются. В такой ситуации предсказанные моделью атомы РНК образуют визуально не разделимое облако точек, а так как ионов магния в структурах РНК достаточно много ( $\approx$  7 ионов магния на 107 нуклеотидов) и в среднем 1 ион оказывает влияние на 20-40 атомов молекулы РНК, при этом ионы металлов распределены неравномерно по всей РНК структуре, они имеют тенденцию группироваться таким образом, чтобы цепочка закручивалась вокруг них. Таким образом, вышеописанные пересечения встречаются в каждой молекуле (за редким исключением), и зачастую, облако предсказанных атомов неразрывно проходит через всю структуру и выглядит достаточно однородным. Например, такую ситуацию можно наблюдать во всех выше приведенных структурах (Рис. 5.1 1hr2).

Очевидно, что данная проблема сильно снижает эффективность работы данного сервиса, потому что одна из главных его задач – помогать ученым-экспериментаторам определять местонахождение ионов магния, обращая их внимание на определенные локации – места, в которых модель отмечает наибольшее влияние. А в связи с перекрытием областей влияния ионов магния очень трудно среди выделенных зон на глаз выделить именно те участки, которые являются наиболее близкими к центрам ионов и которые позволяют вычислить их предположительные координаты.

Поэтому возникает необходимость хотя бы частично автоматизировать процесс выявления наиболее значимых участков среди предсказаний модели. Для этого нужно создать такой инструмент, который бы позволял разделять предсказанные атомы на группы в зависимости от того, какой ион магния оказывает на них наибольшее влияние. Таким образом, в идеале у нас должно получиться столько же групп, сколько ионов магния находится в структуре. А далее логично предположить, что в центрах полученных групп расположены искомые ионы, таким образом вычисляя геометрический центр каждой группы, будет получена аппроксимация координат ионов магния.

Стоит отметить, что полученные координаты следует воспринимать именно как приближение к реальным координатам, поскольку влияние ионов магния рассматривается на довольно большом радиусе (7 Å), в связи с чем могут возникать ситуации, например, когда ион будет находиться за пределами молекулы и у нас просто физически будет не хватать данных, чтобы точно оценить на каком именно расстоянии (до 7 Å) он находится. Также трудно определяемыми являются случаи, когда ионы магния расположены вплотную или почти вплотную, образуя таким

образом целые комплексы. Множества атомов, на которые оказывают влияние ионы из одного такого комплекса, будут практически идентичны, поэтому предложенная аппроксимация будет отождествлять такой комплекс с одним ионом магния.

Математически, постановка данной задачи звучит так: дано множество точек в пространстве, необходимо разделить это множество на  $k$  групп. По сути, это классическая задача машинного обучения без учителя – кластеризация. Существует два основных и наиболее распространенных алгоритма кластеризации:  $k$ -means и DBSCAN.

#### Принцип работы $k$ -means.

Случайным образом или вероятностно выбирается  $k$  точек из предоставленного набора (где число  $k$  задается пользователем), эти точки и есть начальные центры классов. Далее для оставшихся точек вычисляются расстояния до каждой из  $k$  выбранных точек. Считается, что точка принадлежит классу  $i$ , если расстояние от точки до центра класса  $i$  – наименьшее из всех расстояний для этой точки. Таким образом все точки плоскости разделяются на группы, в каждой группе пересчитывается ее центр. И снова все повторяется: деление точек на классы, пересчитывание центров и т.д. до сходимости.

#### Принцип работы DBSCAN.

Идея этого алгоритма состоит в том, что если рядом с точкой на довольно близком расстоянии находится достаточное количество соседей, то эту точку вместе с соседями можно объединить в 1 класс. Количество соседей и расстояние, в пределах которого они рассматриваются, — это гиперпараметры, которые задаются пользователем. В случае, если точка не имеет достаточного числа соседей в рассматриваемом радиусе – она считается отщепленником, то есть не входит ни в одну из групп.

В случае с  $k$ -means необходимо знать заранее количество групп, на которое требуется разбить предоставленную выборку. Применительно к нашей задаче: необходимо знать количество магниев, которые находятся в структуре РНК. Однако, в нашем случае это количество неизвестно: число магниев в разных структурах может быть абсолютно разным и это никак не связано ни с длиной цепочки ни с какими-либо еще параметрами. Также невозможно судить о числе магниев и по количеству предсказанных моделью атомов, так как число атомов, которые могут находиться вблизи одного магния могут свободно варьироваться от 10 и до 60, в зависимости от расположения иона и пространственного расположения цепочки. Подбирать число центров «на глазок» равносильно ситуации до кластеризации, так как не зная реального расположения ионов магния невозможно оценить качество подбора этого параметра.

В отличие от  $k$ -means, DBSCAN определяет количество центров автоматически. Однако, ввиду того, что влияние относительно близко расположенных ионов магния на атомы, создает некоторое однородное облако точек, в большинстве DBSCAN случаев не сможет выделить в этом облаке кластеры, ввиду равноудаленности точек друг от друга. Да и процесс подбора параметров в данном алгоритме производить гораздо труднее.

Таким образом становится понятно, что для решения данной задачи наиболее подходящим алгоритмом является  $k$ -means, однако для хорошей аппроксимации необходимо научиться оценивать количество магниев (центров классов), с учетом имеющейся информации о предсказанных атомах.

Будем идти от обратного: научимся оценивать качество кластеризации для заданного  $k$ . По окончанию кластеризации алгоритм нам выдает координаты центров кластеров, то есть предполагаемые координаты местоположения ионов. Зная эти координаты, мы можем посмотреть на атомы РНК, которые находятся в радиусе  $7\text{\AA}$  от этих центров, и сравнить их с теми предсказанными атомами, которые предлагает наша модель. Тогда мы можем оценить общее покрытие предсказанных атомов нашей кластеризацией, то есть вычислим долю предсказанных атомов, которые охватывает наша кластеризация. Это покрытие и будет нашей оценкой качества кластеризации.

Теперь, когда для любого количества центров мы можем вычислить покрытие, мы можем автоматизировать подбор количества центров для кластеризации: начиная с одного центра будем постепенно увеличивать их количество и вычислять покрытие предсказанных атомов до тех пор, пока эта оценка не достигнет некоторого порога. Понятно, что начиная с некоторого момента покрытие будет равно 1 и продолжая увеличивать число кластеров, наша молекула станет перенасыщенной ионами магния. То есть мы очевидно получим большее число магниев, чем есть на самом деле. И напротив, в ситуациях с довольно низким покрытием, мы будем иметь на выходе недостаток магния. Поэтому необходим некоторый порог покрытия, при достижении которого мы перестанем увеличивать число кластеров, и это число кластеров будет считаться оптимальным.

Какой порог следует установить? Очевидный порог, который приходит в голову, - 1, то есть ситуация при которой мы достигаем полного покрытия. Однако, следует понимать, что процесс кластеризации, все же, дает нам некоторое приближение координат ионов, и допускает незначительные смещения, которые в свою очередь, приводят к тому, что в окружение центра радиусом  $7\text{\AA}$  будут попадать не все предсказанные атомы, а также некоторое число посторонних атомов. Поэтому при единичном пороге мы будем получать ситуацию перенасыщения молекулы и поэтому выбор порога является некоторой эвристикой.

Разумеется, выбранную стратегию аппроксимации координат ионов магния сначала стоит опробовать на реальных данных. То есть вместо предсказанных моделью атомов, будем использовать уже размеченные атомы (для которых целевой признак  $mg$  равен 1). На таких данных предложенная аппроксимация работает очень хорошо. Экспериментально было выведено, что оптимальный порог для покрытия равен 0.9. При такой доле покрытия количество центров кластеров может отличаться от реального на 1-2 она, зато эти центры расположены достаточно близко к реальным ионам. А ситуации, в которых некоторые центры получается немного больше или меньше, в основном соответствуют случаям, когда ионы магния расположены слишком близко друг к другу.

В Таблица 5.1 для структур из тренировочной выборки приведены средние расстояния между координатами предсказанных ионов магния и реальным расположением ионов в порядке возрастания. А также для сравнения показано, какое количество ионов на самом деле содержится в данной структуре РНК и какое количество центров было получено в ходе кластеризации. Расстояния между ионами указаны в ангстремах. С учетом того, что диаметр иона магния составляет  $1.46\text{\AA}$ , можно заключить, что наши приближения в среднем отличаются примерно на 2-3 иона магния.

Таблица 5.1

<b>pdb</b>	<b>5aox</b>	<b>5nzd</b>	<b>1gid</b>	<b>1i9v</b>	<b>462d</b>	<b>2zzm</b>	<b>1hr2</b>	<b>4r4v</b>	<b>4z4d</b>	<b>1hc8</b>	<b>357d</b>
<b>Среднее расстояние</b>	2.080	2.996	3.005	3.258	3.380	3.381	3.415	3.601	3.763	3.974	4.358
<b>Кол-во ионов (реальность)</b>	4	9	24	3	8	5	42	9	3	10	8
<b>Кол-во ионов (предсказание)</b>	3	8	22	3	6	5	44	8	3	9	5

Для демонстрации полученных результатов ниже построены визуализации для трех цепочек с разным качеством аппроксимации (судя по среднему расстоянию): 5nzd, 462d и 1hc8. Также для наглядности проведены кластеризации с разным порогом покрытия, чтобы увидеть, как ведет себя k-means при разных k (Рис. 5.5). Стоит отметить, что при снижении порога заметно уменьшается количество центров в то время, как среднее расстояние между центрами кластеров до ближайших ионов магния почти не меняется. В принципе, аналогичная ситуация наблюдается и при повышении порога до единицы, хотя зачастую можно увидеть явление децентрализации (например, в молекуле 5nzd), где верно приближенный центр при доле покрытия 0.9 аппроксимируется 3 центрами в некоторое окрестности при доле покрытия 1.

Таким образом, мы выяснили, что на «идеальных», то есть правильно размеченных, данных, предложенный способ аппроксимации центров ионов работает довольно точно. Однако, нас больше интересует, как данный алгоритм справится с зашумленными данными – теми предсказаниями, которые мы получаем от нашей модели. Повторим весь процесс кластеризации на предсказаниях модели, порог покрытия оставим тот же, что был выбран и в случае размеченных данных. В Таблица 5.2 приведены средние расстояния, аналогичные тем, что вычислялись для кластеризации на заранее размеченных данных (Таблица 5.1), а на Рис. 5.6 приведены визуализации для тех же отобранных цепочек.

Качество кластеризации на данных, полученных от модели RandomForest заметно хуже. Здесь наименьшее среднее расстояние между предполагаемыми и реальными центрами ионов больше максимального расстояния, полученного в прошлой кластеризации, и количества аппроксимированных центров значительно возросли. Разумеется, это была ожидаемая ситуация, поскольку известно, что выстроенная модель, предсказывающая атомы, на расстоянии 7 Å от которых находятся ионы магния, приемлемо работает далеко не со всеми структурами и в целом имеет тенденцию предсказывать избыточное число атомов. А алгоритм кластеризации k-means естественным образом достаточно чувствителен к разного рода отклонениям. Из-за этого даже для молекулы 462d, для которой модель выдает предсказания довольно высокого качества, центры ионов получились немного смешенными.

Таблица 5.2

<b>pdb</b>	<b>462d</b>	<b>1i9v</b>	<b>1gid</b>	<b>1hr2</b>	<b>5nzd</b>	<b>429d</b>	<b>4z4d</b>	<b>1hc8</b>	<b>4r4v</b>	<b>2zzm</b>	<b>5aox</b>
<b>Среднее расстояние</b>	4.68	4.78	5.30	5.62	5.94	6.65	6.78	7.1	8.78	9.79	15.69
<b>Кол-во ионов (реальность)</b>	8	3	24	42	9	2	3	10	9	5	4
<b>Кол-во ионов (предсказание)</b>	7	15	65	60	15	8	3	13	37	15	17

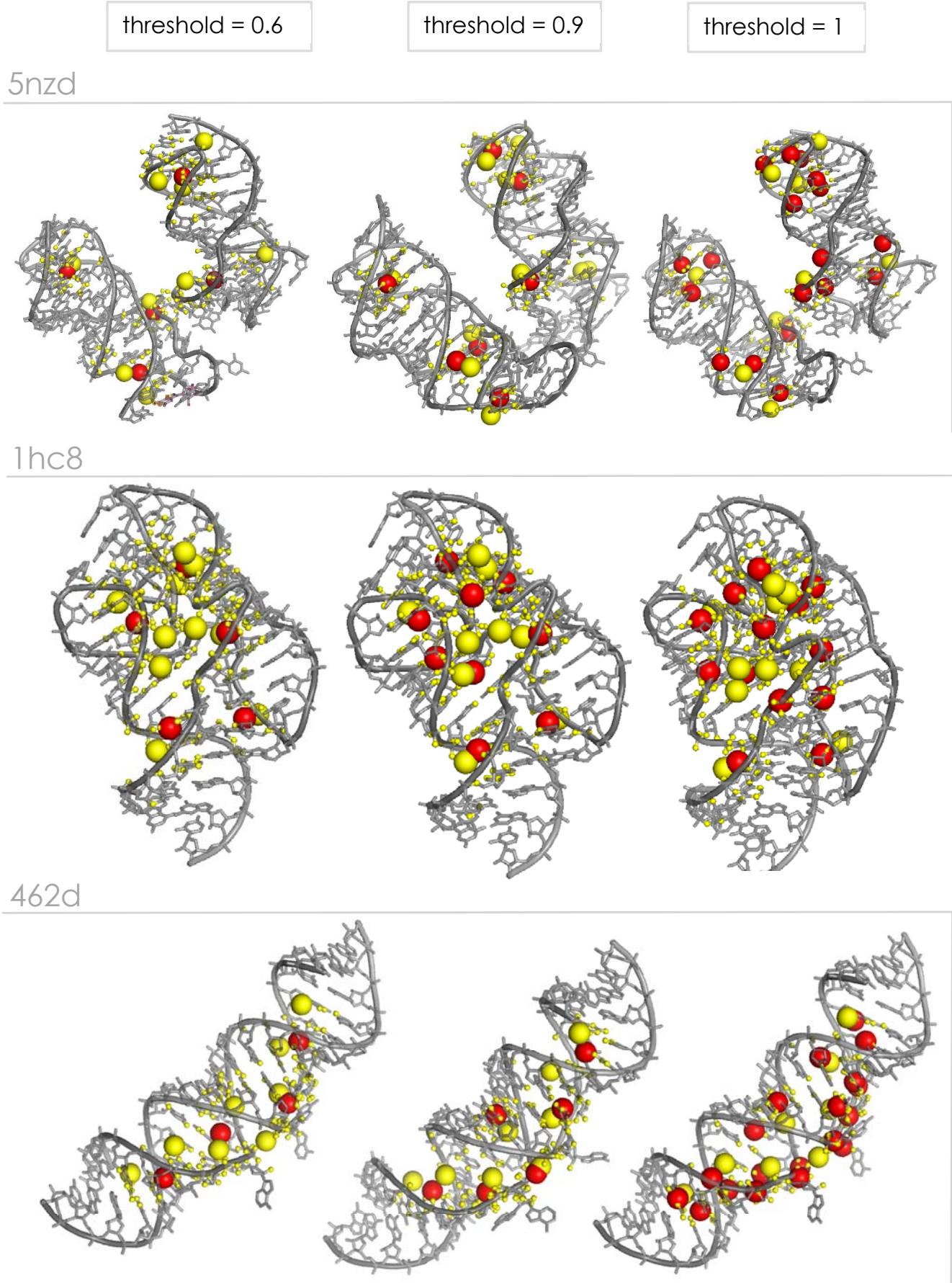


Рис. 5.5

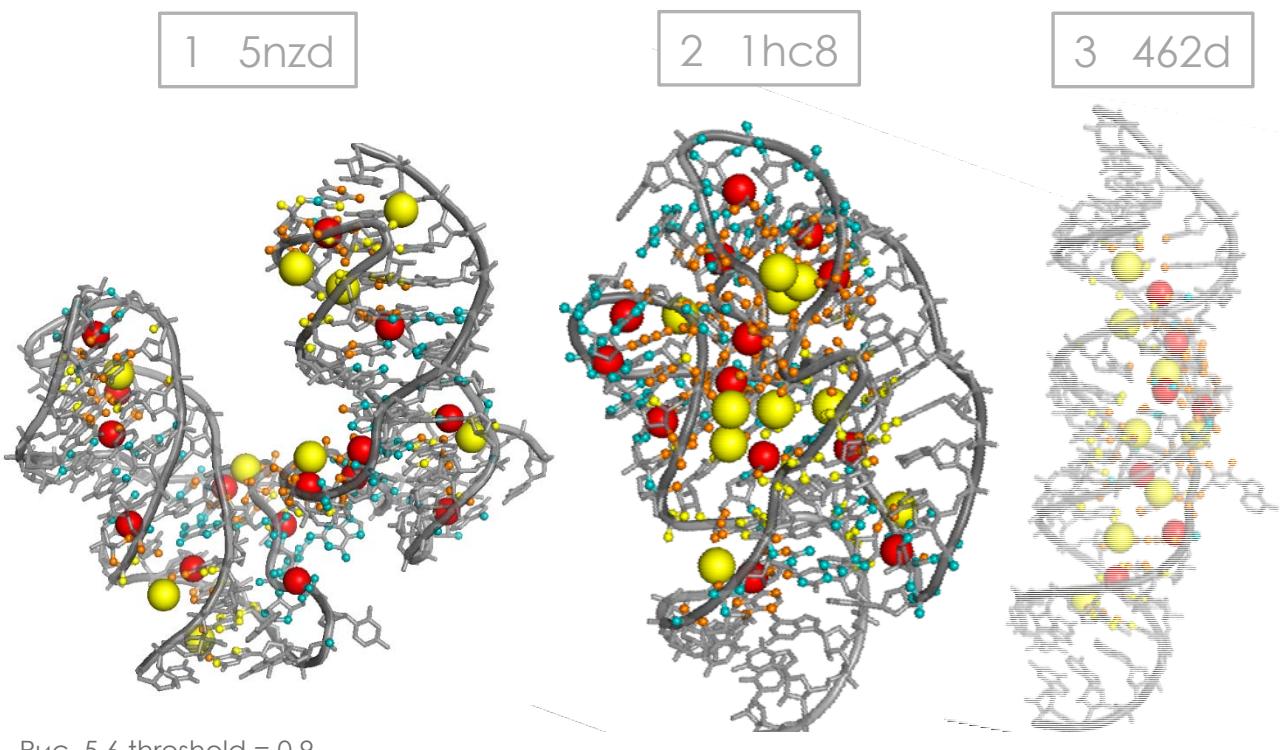


Рис. 5.6 threshold = 0.9

Учитывая избыточность предсказаний, стоит понизить порог покрытия. Эксперименты показывают, что, снизив порог хотя бы на 0.05, аппроксимация центров ионов получается точнее. Однако выбор доли покрытия, в конечном счете остается за пользователем. При подборе этого порога стоит опираться в первую очередь на то, как предполагаемые центры сопоставляются с выделенными атомами, а также на их плотность распределения.

Заметим, что в структурах, предсказания которых содержат минимальное число посторонних атомов (например, 462d), модель в принципе предсказывает атомы вокруг почти всех ионов магния, но на расстоянии меньше 7 Å. Это приводит к мысли о том, что если кластеризовать предсказанные атомы с избыточным числом классов, тогда центры этих классов будут децентрализовывать реальные расположения магниев + мы получим также некоторые группы центров в местах избыточного предсказания ионов. В таком случае мы сможем кластеризовать полученные центры кластеров, объединяя наиболее близко расположенные центры в одну группу. Этот процесс можно реализовать с помощью алгоритма DBSCAN, таким образом от сильно избыточного числа предсказанных ионов, мы сможем приблизиться к реальному числу.

Однако метод DBSCAN требует задания двух гиперпараметров: числа соседей и радиуса поиска соседей. И если количество соседей будет зависеть от того, во сколько раз больше мы зададим классов для кластеризации, то выбор радиуса поиска соседей опять остается на усмотрение пользователя. Правда здесь можно опираться на гистограмму распределения расстояний между центрами, которые мы хотим кластеризовать.

В качестве примера, проделаем эту процедуру двойной кластеризации со структурой 462d. Так, вполне достаточным будет увеличить количество классов вдвое. В таком случае, мы будем рассматривать пары точек и объединять их в один класс, если они находятся на расстоянии  $d$  друг от друга. Это расстояние мы определим по гистограмме (Рис. 5.7), приблизительно, оно должно быть в районе 7 Å.

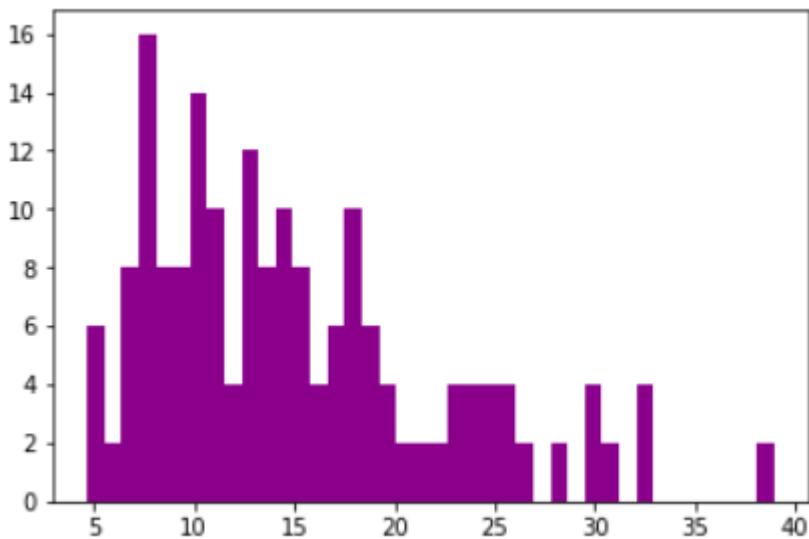


Рис. 5.7

На этой гистограмме пик расстояний между центрами приходится на  $7.5 \text{ \AA}$ . Таким образом, имеет смысл объединять все точки, которые находятся на расстоянии  $7.5 \text{ \AA}$  и меньше друг от друга. К слову сказать, количество центров для первой кластеризации не стоит выбирать слишком большим, иначе мы вернемся к изначальной проблеме однородного облака точек, с которой DBSCAN не справляется. В то же время центров должно быть и не слишком мало: в идеале, минимально расстояние между центрами не должно быть меньше  $4 \text{ \AA}$  и не превышать  $7 \text{ \AA}$ . Возвращаясь к нашему примеру, для молекулы 462d получилась следующая аппроксимация (Рис. 0.1) (среднее расстояние между приближенными и реальными ионами =  $4.09$ , ионов предсказано: 7 из 8).

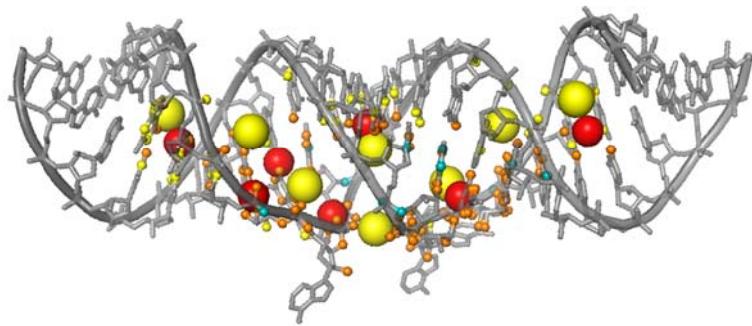


Рис. 0.1

Таким образом, основная суть метода двойной кластеризации в том, что он централизует предсказанные координаты, и за счет этого, компенсирует смещения, вызванные недопредсказанием модели. Поэтому такой подход может быть полезен только в работе с не зашумленными данными. Однако на данный момент не так много структур, для которых RandomForest выдает предсказания с малым числом посторонних атомов, а также пока что невозможно определить, какие именно структуры будут хорошо предсказываться данной моделью.

## 6. Заключение

Итак, в данной работе разработана модель для прогнозирования того, рядом с какими атомами данной молекулы РНК располагаются ионы магния. Рядом в данном случае означает расстояние до 7 Å. Также был разработан алгоритм, аппроксимирующий полученные данные с целью получения приближенных значений координат ионов. В следствие неидеальной работы модели, полученные приближения осмысленно рассматривать только в совокупности с предсказанными атомами.

По результатам исследований можно сделать вывод, что на данный момент не существуют сервиса, который бы позволял определять местонахождение ионов магния с высокой точностью для всех структур. Однако тот факт, что существующие сервисы, в том числе и представленный в данной работе алгоритм, в некоторых случаях довольно точно оценивают координаты/влияние иона магния, говорит о том, что задача распознавания связей ионов металлов со структурами РНК может быть решена для определённых групп структур. Также, ожидается, что качество предсказаний возрастет с ростом количества качественных экспериментальных данных.

Все сопроводительные коды этой работы можно найти на [github](#):

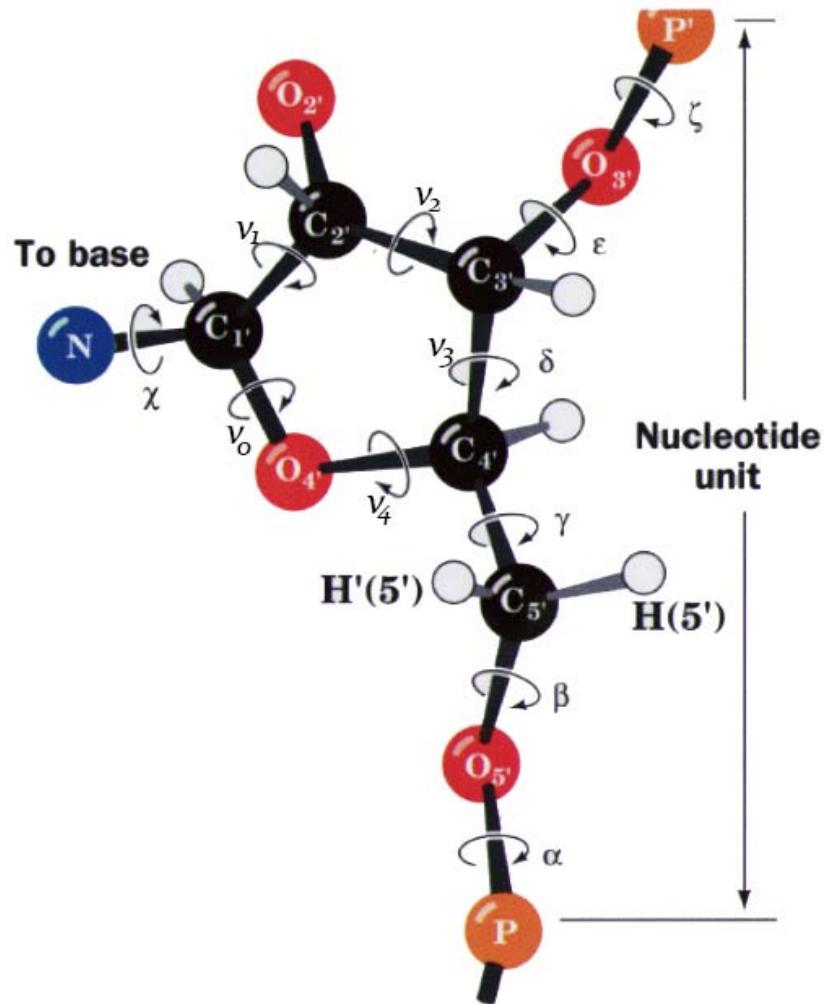
<https://pollytikhonova.github.io/coursework/>

## 7. Список литературы

1. J. S. Mattick, I. V. Makunin, «Non-coding RNA», *Human Molecular Genetics*, vol. 15, 2006.
2. V. K. Misra, D. E. Draper, «On the role of magnesium ions in RNA stability», *Biopolymers*, vol. 48, pp. 113-135, 1998.
3. D. R. Banatao, R. B. Altman, T. E. Klein, «Microenvironment analysis and identification of magnesium binding sites in RNA», *Nucleic Acids Research*, vol. 31, № 15, pp. 4450-4460, 2003.
4. D. E. Draper, «A guide to ions and RNA structure», *RNA*, vol. 10, № 3, pp. 335-343, 2004.
5. I. Turel, J. Kljun, «Interactions of Metal Ions with DNA, Its Constituents and Derivatives, which may be Relevant for Anticancer Research», *Current Topics in Medicinal Chemistry*, vol. 11, № 21, pp. 2661-2687, 2011.
6. R. Römer, R. Hach, «tRNA Conformation and Magnesium Binding», *FEBS Journal*, vol. 55, № 1, pp. 271-284, 1975.
7. A. Stein, D. M. Crothers, «Equilibrium binding of magnesium(II) by Escherichia coli tRNA<sub>fMet</sub>», *Biochemistry*, vol. 15, № 1, pp. 157-160, 1976.
8. T. Hermann, E. Westhof, «Exploration of metal ion binding sites in RNA folds by Brownian-dynamics simulations», *Structure*, vol. 6, № 10, pp. 1303-1314, 1998.
9. V. K. Misra, D. E. Draper, «Mg 2+ binding to tRNA revisited: the nonlinear poisson-boltzmann model», *Journal of Molecular Biology*, vol. 299, № 3, pp. 813-825, 2000.
10. «<https://omictools.com/protein-metal-site-prediction-category>», OMICtools, 2015. [В Интернете]. Available: <https://omictools.com/protein-metal-site-prediction-category>.
11. A. Philips, K. Milanowska, G. Lach, M. Boniecki, K. Rother, J. M. Bujnicki, «MetalionRNA: computational predictor of metal-binding sites in RNA structures», *Bioinformatics*, vol. 28, № 2, pp. 198-205, 2012.
12. E. Y. V. Baulin, D. Khachko, S. Spirin, M. Roytberg, «URS DataBase: universe of RNA structures and their motifs», *Database (Oxford)*, 2016.
13. R. P. D. Bank, «RCSB Protein Data Bank - RCSB PDB», 2000. [В Интернете]. Available: <http://www.rcsb.org/pdb/home/>.
14. N. B. Leontis, C. L. Zirbel, «Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking», 2012. [В Интернете]. Available: [https://link.springer.com/content/pdf/10.1007/978-3-642-25740-7\\_13.pdf](https://link.springer.com/content/pdf/10.1007/978-3-642-25740-7_13.pdf).
15. П. Тихонова, «Анализ связывания ионов магния с РНК», 2017. [В Интернете]. Доступно: [https://lms.hse.ru/ap\\_service.php?getwork&guid=7B067524-578F-49B3-AE35-E4C7A3DE2440](https://lms.hse.ru/ap_service.php?getwork&guid=7B067524-578F-49B3-AE35-E4C7A3DE2440)
16. B. G. Beltchev, M. . Yaneva, D. Z. Staynov, «Thermal Melting Curves of tRNAPhe from Yeast Lacking Different Numbers of Nucleotides from the 3'-End», *FEBS Journal*, vol. 64, № 2, pp. 507-510, 1976.
17. V. a. Voet, «Biochemistry», John Wiley & Sons, 1990.

## Приложение А

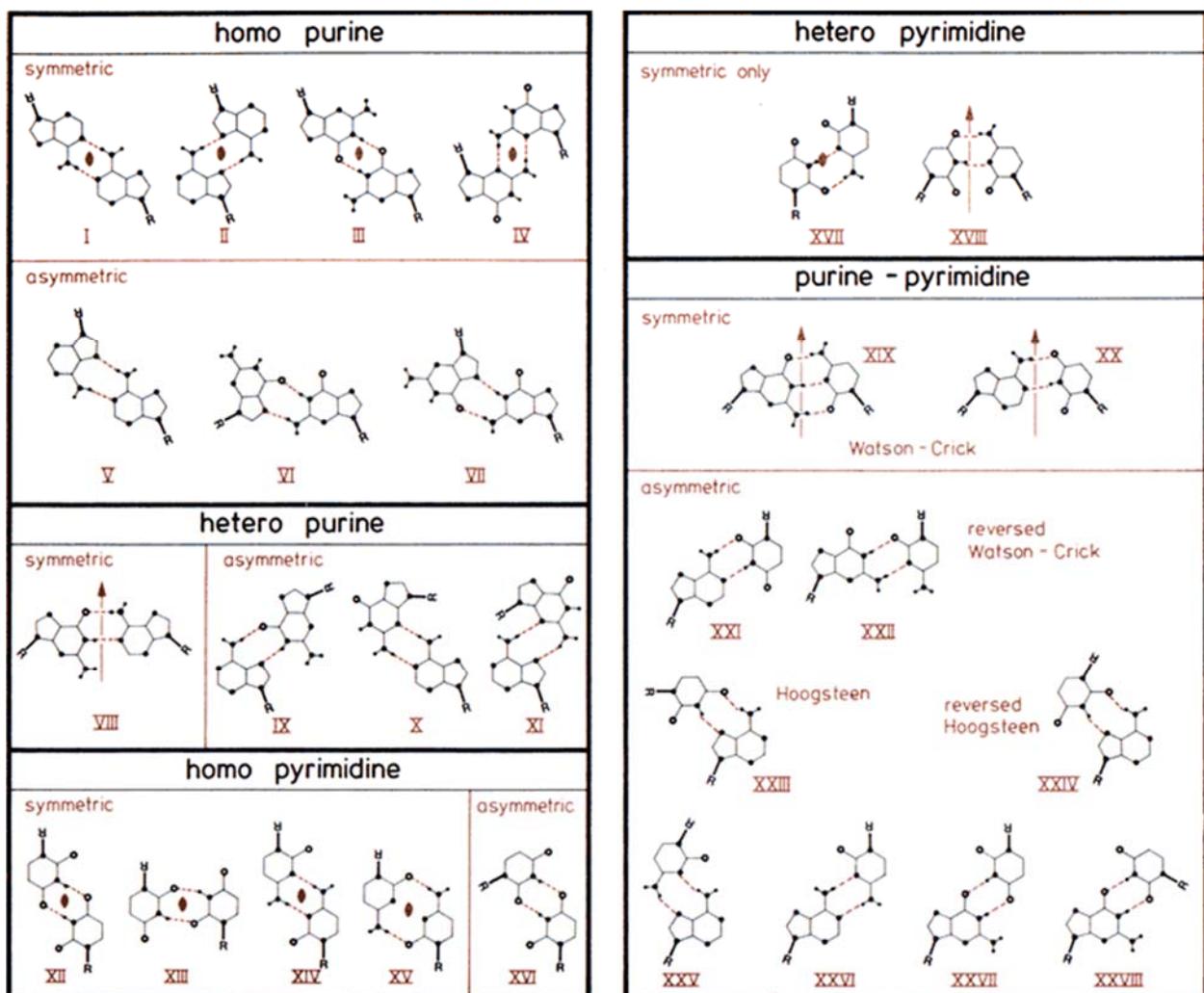
### Торсионные углы



[Источник: D. Voet, J. G. Voet, "Biochemistry", John Wiley and Sons, New York. 1990]

## Приложение В

### Классификация спариваний оснований по Сэнгеру

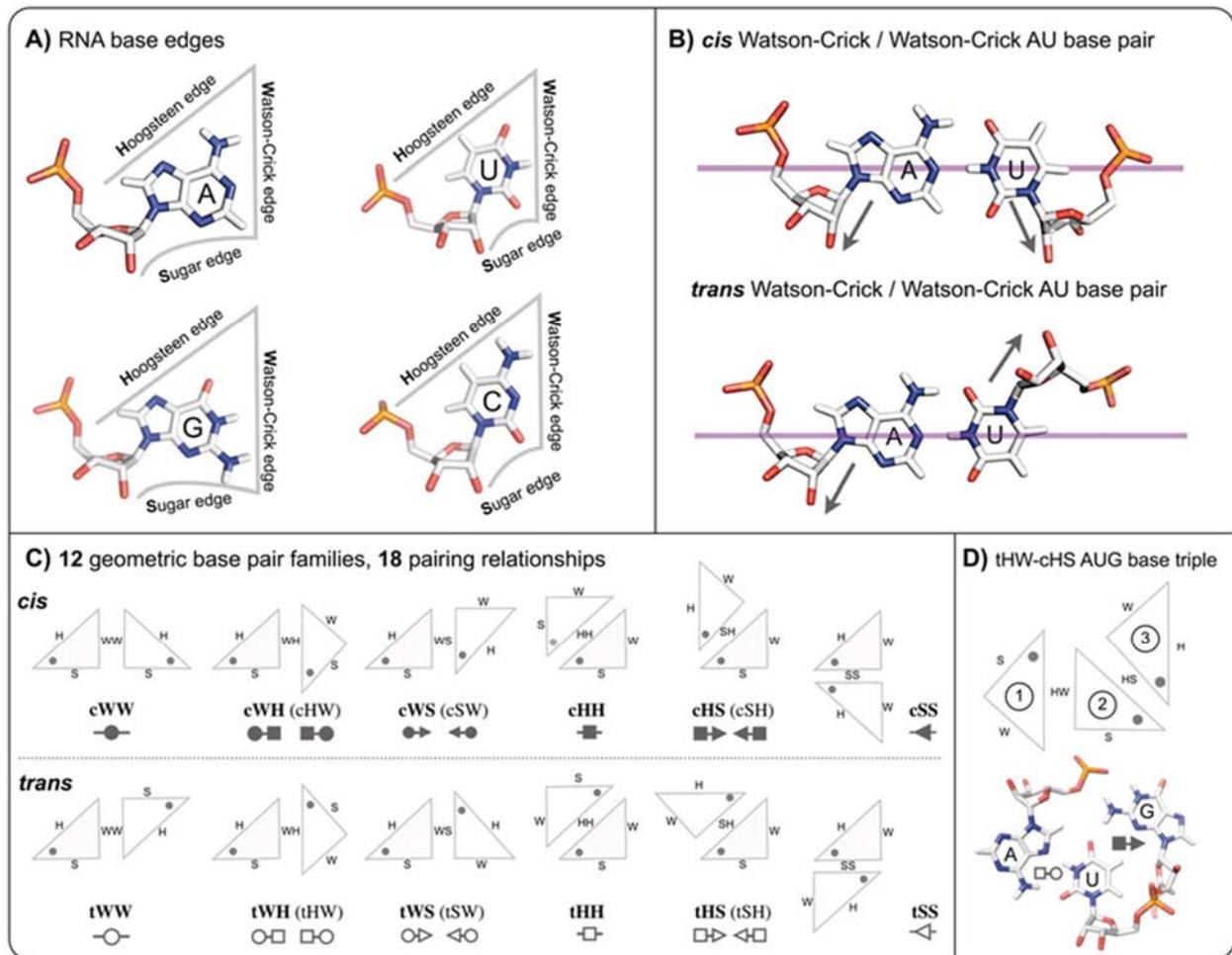


[Источник: <http://x3dna.org/highlights/reverse-watson-crick-base-pairs>]

# Приложение С

## Классификация спариваний оснований

### Леонтиса-Вестхофа



[A. Almakarem, A. Petrov, J. Stombaugh, C. Zirbel, N. Leontis, "Comprehensive survey and geometric classification of base triples in RNA structures", *Nucleic acids research*, 2011]