**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**Кафедра биомедицинской информатики**

КАМЕНКО

Яна Александровна

**АЛГОРИТМ ДЛЯ ПРЕДСКАЗАНИЯ СТРУКТУРЫ БЕЛКОВОГО КОМПЛЕКСА НА ОСНОВЕ СТРУКТУРНЫХ ШАБЛОНОВ**

Дипломная работа

Научный руководитель:

старший преподаватель кафедры БМИ,

А.Ю. Хадарович

Допущена к защите

“\_\_\_” \_\_\_\_\_\_\_\_\_\_\_ 2020 г

Зав. кафедрой биомедицинской информатики

кандидат физико-математических наук,

доцент Ю.Л. Орлович

Минск, 2020

**ОГЛАВЛЕНИЕ**

ВВЕДЕНИЕ ………………………………………………………………........ 3

ГЛАВА 1 ПОСТАНОВКА ЗАДАЧИ ………………………………............... 7

1.1 Цель ………………………………………………………………………... 8

1.2 Набор входных данных ……………………………………....................... 8

ГЛАВА 2 АЛГОРИТМ СТРУКТУРНОГО ВЫРАВНИВАНИЯ …………... 10

2.1 Среднеквадратичное отклонение ………………………………...…........ 10

2.2 Оценка шаблонного моделирования …………………………………….. 11

2.3. Алгоритм TM-align ………………………………………………………. 12

2.4 Визуализация молекул …………………………………………………… 14

ГЛАВА 3 ПРИМЕНЕНИЕ АЛГОРИТМА ДЛЯ ВХОДНЫХ НАБОРОВ ДАННЫХ ……………………………………………………………………… 15

3.1 Алгоритм TM-align ………………………………………………….…….. 16

3.2 Таргет-белок ….........................................................................……………. 16

3.3 Оценка шаблонного моделирования ……………………………...…........ 17

3.4 Топ-10 …………………………………………….………………………… 18

3.5 Визуализация молекул …………………………………..………………… 21

3.6 Среднеквадратичное отклонение ……………………………………….… 28

ГЛАВА 4 УСОВЕРШЕНСТВОВАНИЕ АЛГОРИТМА …………………….. 29

4.1 Обработка входных данных ………………………………………………. 29

4.2 Текущий алгоритм ……………………………………………………….. 30

4.3 Улучшение алгоритма ……………………………………………………. 30

4.4 Типы графиков ……………………………………………………………. 31

4.5 Графики для изначального алгоритма …………………………………... 32

4.6 Графики для улучшенного алгоритма …………………..………………. 36

4.7 Сравнение алгоритмов …………………………………………..……….. 38

ГЛАВА 5 МОДЕЛИРОВАНИЕ МЕТАБОЛИЧЕСКОГО ПУТИ ………...... 39

5.1 Метаболический путь ………………………………………………….…. 39

5.2 База данных UniProt ……………………………………………………… 39

5.3 Проект KEGG ……………………………………………………………... 40

5.4 Применение алгоритма для метаболического пути ……………………. 41

ЗАКЛЮЧЕНИЕ …………………………………………………………...….. 44

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ……………………...…. 45

**РЕФЕРАТ**

Дипломная работа, 45 страниц, 31 рисунок, 6 таблицы, 5 формул, 8 источников.

**Ключевые слова:** БЕЛКОВЫЕ КОМПЛЕКСЫ, ШАБЛОНЫ, ПОЛНЫЕ СТРУКТУРЫ, ИНТЕРФЕЙСЫ, ДОКИНГ, СТРУКТУРНОЕ ВЫРАВНИВАНИЕ.

**Объект исследования:** белковые комплексы, шаблоны белковых комплексов.

**Цель работы:** разработка алгоритма для предсказания белкового комплекса на основе структурных шаблонов.

**Результат:** изучена научная литература по теме докинга, рассмотрен алгоритм структурного выравнивания TM-align, определены входные данные и для этих данных применен алгоритм TM-align, результаты докинга визуализированы с помощью программы pyMOL, доработан алгоритм, оценены его точность, использован улучшенный алгоритм для моделирования метаболического пути

**Область применения:** моделирование белковых комплексов.

**РЭФЕРАТ**

Дыпломная праца, 45 старонак, 31 малюнак, 6 табліцы, 5 формул, 8 крыніц.

**Ключавыя словы:** БЯЛКОВЫЯ КОМПЛЕКСЫ, ШАБЛОНЫ, ПОЎНЫЯ СТРУКТУРЫ, ІНТЭРФЕЙСЫ, ДОКІНГ, СТРУКТУРНАЕ ВЫРАЎНАННЕ.

**Аб’ект даследавання:** бялковыя комплексы, шаблоны бялковых комплексаў.

**Мэта работы:** распрацоўка алгарытму для прадказання бялковага комплексу на грунце структурных шаблонаў.

**Вынік:** вывучана навуковая літаратура па тэме докінга, разгледжаны алгарытм структурнага выраўнання TM-align, вызначаны ўваходныя дадзеныя і для гэтых дадзеных ужыты алгарытм TM-align, вынікі докінга візуалізаваны з дапамогай праграмы pyMOL, дапрацаваны алгарытм, ацэнены яго дакладнасць, скарыстаны палепшаны алгарытм для мадэлявання метабалічнага шляху.

**Вобласць прымянення:** мадэляванне бялковых комплексаў.

**ABSTRACT**

Diploma thesis, 45 pages, 31 figures, 6 tables, 5 formulas, 8 sources.

**Keywords:** PROTEIN COMPLEXES, TEMPLATES, FULL STRUCTURES, INTERFACES, DOCKING, STRUCTURAL ALIGNMENT.

**Object of research:** protein complexes, templates of protein complexes.

**Objective:** Development of an algorithm for prediction of a protein complex based on structural templates.

**The result:** the scientific literature on the topic of docking has been studied, the algorithm of structural alignment TM-align has been considered, input data have been defined and the algorithm TM-align has been applied for this data, the results of docking have been visualized with the help of the pyMOL program, the algorithm has been refined, its accuracy has been evaluated, an improved algorithm has been used to simulate the metabolic pathway.

**The scope:** modeling protein complexes.

**ВВЕДЕНИЕ**

Современная биология имеет дело с колоссальными объемами данных, откуда возникает необходимость структурирования информации и оптимизации процессов подбора исходной данных для дальнейших исследований. Это стало причиной создания такой междисциплинарной области, как биоинформатика. Биоинформатика объединяет с одной стороны молекулярную биологию и генетику, с другой стороны – компьютерные науки, математику и статистику. Специалисты в биоинформатике занимаются изучением и разработкой компьютерных методов, направленных на получение, анализ, хранение, организацию и визуализацию биологических данных.

Одним из главных объектов изучения биоинформатики является белок. Выделяют 4 уровня структурной организации белка: первичная, вторичная, третичная и четвертичная. В свой дипломной работе я занималась изучением третичной структуры, а конкретнее, молекулярным докингом, позволяющим предсказать наиболее выгодную для образования устойчивого комплекса ориентацию и конформацию одной молекулы в центре связывания другой.

Таким образом, целью моей дипломной работы является разработка алгоритма для предсказания белкового комплекса на основе структурных шаблонов.

Главными целями создания такого алгоритма являются ускорение и уменьшение стоимости разработки лекарственных средств в сфере биотехнологий, занимающейся поиском препаратов для точечной терапии тяжелых заболеваний.

**ГЛАВА 1. ПОСТАНОВКА ЗАДАЧИ**

**1.1 Цель**

Цель дипломной работы – разработать алгоритм для предсказания белкового комплекса на основе структурных шаблонов. Для этого нужно изучить научную литературу по теме докинга, рассмотреть алгоритм структурного выравнивания TM-align, определить входные данные и для этих данных применить алгоритм TM-align, результаты докинга визуализировать с помощью программы pyMOL, доработать алгоритм, оценить его точность, использовать улучшенный алгоритм для моделирования метаболического пути.

**1.2 Набор входных данных**

Общедоступный ресурс Dockground [1] предоставляет данные для улучшения понимания межбелковых взаимодействий и помогает в разработке более совершенных инструментов для структурного моделирования белковых комплексов, таких как алгоритмы докинга и оценочные функции.

В качестве входных данных я использовала наборы данных под названиями “Full structures v1.1”, где шаблонами являются полные структуры, и “Interfaces v1.1”, где шаблонами являются интерфейсы – взаимодействующие участки белкового комплекса. Наборы данных примечательны тем, что исходный набор всех связанных гетеро- и гомодимеров из Dockground сокращен с использованием следующих требований:

1) рентгеновские структуры с разрешением больше или равно 3,5 ;

2) количество остатков на границе раздела в каждой цепи больше или равно 10;

3) наличие как минимум трех регулярных элементов вторичной структуры (альфа-спирали и / или бета-нити) в каждой субъединице (только для набора данных “Full structures v1.1”).

Применение этих критериев привело к 12134 избыточным комплексам. Структурная избыточность устранена с помощью MM-align алгоритма и с использованием алгоритма кластеризации графов Хартува и Шамирома. Набор данных “Full structures v1.1” содержит 4950 шаблонов полных структур без избыточности, набор данных “Interfaces v1.1” - 5936 шаблонов интерфейсов без избыточности.

Следующие обозначения используются для наименования PDB-файлов:

iXXXXM1CH1M2CH2\_N.pdb, где

i – указывает, что шаблон взят из библиотеки белковых интерфейсов (набор данных “Interfaces v1.1”);

XXXX – код PDB;

M1, M2 - серийный номер модели в соответствующем файле Biounit для цепей CH1 и CH2;

CH1, CH2 - идентификаторы цепи для двух взаимодействующих белков;

N - 1 для цепи CH1 и 2 для цепи CH2.

**ГЛАВА 2. АЛГОРИТМ СТРУКТУРНОГО ВЫРАВНИВАНИЯ**

**2.1. Среднеквадратичное отклонение атомных положений**

В биоинформатике среднеквадратичное отклонение атомных положений (или просто среднеквадратичное отклонение, RMSD) является мерой среднего расстояния между атомами (обычно атомами основной цепи) наложенных белков. При изучении конформаций глобулярного белка обычно измеряют сходство трехмерной структуры с помощью среднеквадратичного отклонения атомных координат Cα после оптимальной суперпозиции твердого тела.

Широко используемый способ сравнения структур биомолекул или твердых тел состоит в том, чтобы перемещать и вращать одну структуру относительно другой, чтобы минимизировать среднеквадратичное отклонение. Среднеквадратичное отклонение вычисляется по формуле (1):

, (1)

где - расстояние между атомом и эталонной структурой или среднее положение N эквивалентных атомов. Это часто рассчитывается для тяжелых атомов основной цепи C, N, O и Cα или иногда только для атомов Cα.

Обычно выполняется жесткая суперпозиция, которая минимизирует среднеквадратичное отклонение, и этот минимум возвращается. Для двух наборов точек и среднеквадратичное отклонение определяется по формуле (2):

(2)

Значение среднеквадратичного отклонения выражается в единицах длины. Наиболее часто используемой единицей в структурной биологии является Ангстрем (Å), который равен м.

**2.2. Оценка шаблонного моделирования**

Оценка шаблонного моделирования, или TM-score, является показателем сходства между двумя белковыми структурами с различными третичными структурами [4].

Поскольку среднеквадратичное отклонение взвешивает расстояния между всеми парами остатков одинаково, небольшое количество локальных структурных отклонений может привести к высокому среднеквадратичному отклонению, даже если глобальные топологии сравниваемых структур схожи. Кроме того, среднее среднеквадратичное отклонение случайно связанных белков зависит от длины сравниваемых структур, что делает абсолютную величину среднеквадратичное отклонение бессмысленной. Оценка шаблонного моделирования преодолевает эти проблемы, используя вариацию весового коэффициента Левитта – Герштейна (LG), которая сравнивает пары остатков на меньших расстояниях сильнее, чем пары на больших расстояниях. Следовательно, оценка шаблонного моделирования более чувствительна к глобальной топологии, чем к локальным структурным изменениям.

Оценка шаблонного моделирования указывает разницу между двумя структурами с помощью оценки между , где 1 указывает на идеальное совпадение между двумя структурами (чем выше, тем лучше). Как правило, оценка ниже 0,20 соответствуют случайно выбранным неродственным белкам, в то время как структуры с оценкой выше 0,5 предполагают примерно одинаковую складку. Оценка шаблонного моделирования рассчитывается независимо от длины белка. Оценка шаблонного моделирования вычисляется по формуле (3).

(3)

где и - длины таргет-белка и выровненного участка соответственно, – расстояние между парой остатков, а - шкала расстояний, нормализующая расстояния.

**2.3. Алгоритм TM-align**

Алгоритм TM-align [6] использует только основные -координаты заданных белковых структур.

*Начальное структурное выравнивание*

Используются три вида быстро идентифицируемых начальных выравниваний. Первый тип начального выравнивания получается путем выравнивания вторичных структур двух белков с использованием динамического программирования. Элемент матрицы оценок назначается равным 1 или 0 в зависимости от того, идентичны ли элементы вторничой структуры выровненных остатков. Здесь штраф -1 за открытие пропуска работает лучше всего. Для данного остатка состояние вторичной структуры ( или спираль) назначается на основе координат Cα пяти соседних остатков, то есть -й остаток назначается как , в то время как

(4)

выполняется для всех , которые обозначают расстояние Cα между -м и -ым остатками; в противном случае он назначается спиралью. Окончательное назначение сглаживается путем слияния и удаления синглетных состояний вторичной структуры.

Второй тип начального выравнивания основан на сопоставлении двух структур без пропусков. Для меньшего из двух сравниваемых белков мы выполняем обработку без пропусков для более крупной структуры, но вместо использования среднеквадратичного отклонения в качестве метрики сравнения, как это было сделано в других методах, теперь выбирается выравнивание с лучшей оценкой шаблонного моделирования.

Третье начальное выравнивание также получается с помощью динамического программирования с использованием штрафа -1 за открытие пропуска, но матрица оценок представляет собой комбинацию матрицы оценок вторичных структур и матрицы оценок расстояний, выбранных во втором начальном выравнивании.

*Эвристическая итерация*

Полученные выше начальные выравнивания представлены в эвристическом итеративном алгоритме, который широко использовался при оптимизации NP-сложных структурных выравниваний. В этой процедуре сначала вращаются структуры с помощью матрицы вращения оценки шаблонного моделирования на основе выровненных остатков в начальных выравниваниях. Матрица сходства скора определяется как

(5)

где - расстояние -го остатка в структуре 1 и -го остатка в структуре 2 в суперпозиции оценки шаблонного моделирования; , где - длина меньшего белка. Новое выравнивание может быть получено путем реализации динамического программирования на матрице с оптимальным штрафом за открытие промежутка -0,6. Затем мы снова накладываем структуры на матрицу вращения оценки шаблонного моделирования в соответствии с новым выравниванием и получаем более новое выравнивание путем реализации динамического моделирования с новой оценочной матрицей. Процедура повторяется до тех пор, пока выравнивание не станет стабильным и не будет возвращено выравнивание с наибольшим значением оценки шаблонного моделирования. Из-за непротиворечивости матрицы вращения на основе оценки шаблонного моделирования и показателя сходства динамического программирования выравнивания сходятся очень быстро, и обычно для определения наилучшего выравнивания достаточно 2-3 итераций.

Алгоритм TM-align работает в 4 раза быстрее, чем алгоритм CE, и в 20 раз быстрее алгоритмом DALI и SAL [3]. В среднем полученные структурные выравнивания имеют более высокую точность и охват, чем выравнивания, полученные с помощью других методов.

Алгоритм TM-align также примечателен тем, что он является одним из немногих поддерживаемых алгоритмов – его постоянно совершенствуют. Самая новая версия алгоритма – за 22.08.2019.

**2.4 Визуализация молекул**

PyMOL [5] (рисунок 2.1) — одна из немногих систем молекулярной визуализации с открытым исходным кодом, пригодная для использования в структурной биологии. Данная программа позволяет создавать высококачественные трёхмерные изображения как малых молекул, так и биологических макромолекул, в первую очередь белков. Примерно четверть всех публикуемых в научной литературе изображений структур белков сделана с помощью PyMOL.

Программа PyMOL удобна в использовании. Можно загружать готовые скрипты на языке Python, а также загружать скрипты с расширением .pml. Командная строка PyMOL хранит все сессионные команды, что облегчает работу с молекулами.

В своей дипломной работе я использовала программу pyMOL для визуализации результатов выравниваний алгоритма TM-align.

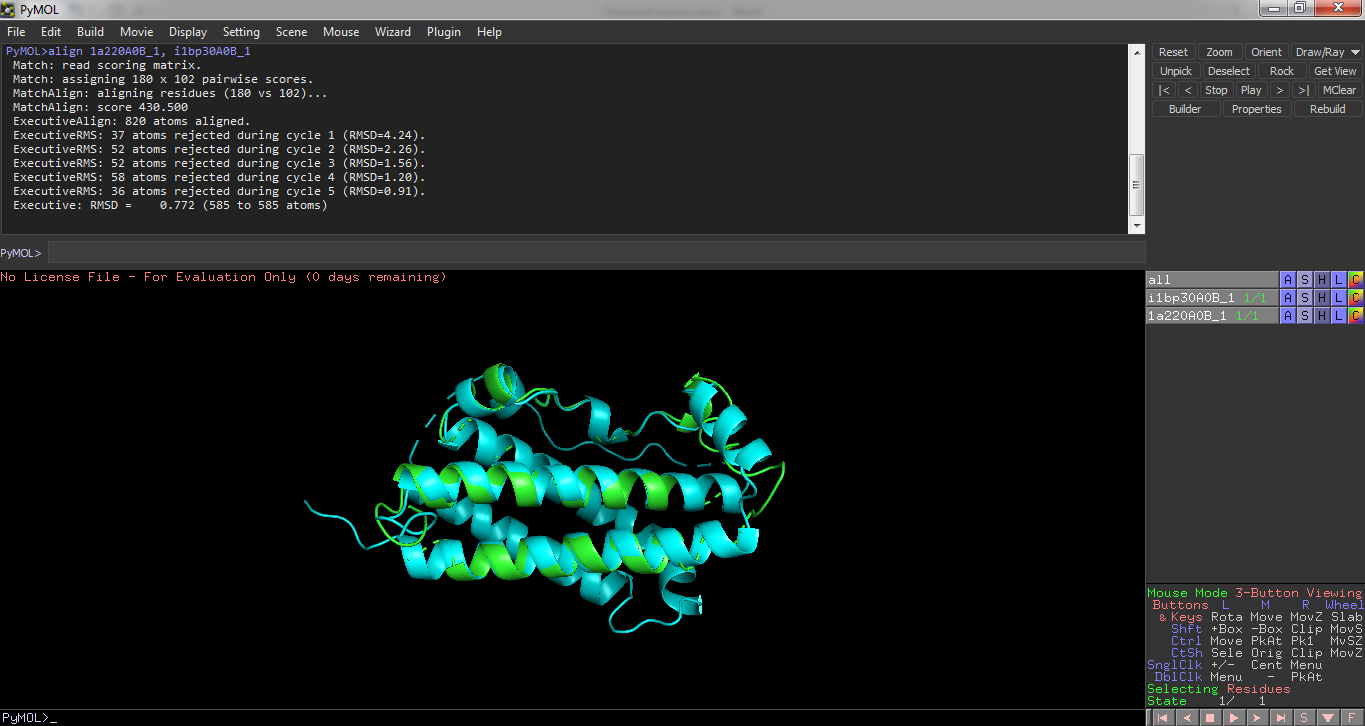


Рисунок 2.1 – Программа pyMOL

**ГЛАВА 3. ПРИМЕНЕНИЕ АЛГОРИТМА TM-ALIGN ДЛЯ ВХОДНЫХ НАБОРОВ ДАННЫХ**

Для ознакомления с алгоритмом TM-align я проделала следующие шаги:

1. Загрузила алгоритм TM-align и входной набор данных;
2. Запустила алгоритм TM-align для любого таргет-белка из набора данных со всеми шаблонами;
3. Обработала отчет, выдаваемый TM-align, выбратла нужную оценку шаблонного моделирования;
4. На основании всех полученных оценок шаблонного моделирования составила топ-10 лучших структурных выравниваний и топ-10 худших выравниваний;
5. Лучшее и худшее выравнивания визуализировала с помощью программы pyMOL;
6. Вычислила среднеквадратичное отклонение для этих выравниваний, сравнила значения среднеквадратичного отклонения со значениями оценки шаблонного моделирования.
   1. **Алгоритм TM-align**

Исходный код алгоритма TM-align [7] я брала на сайте Zhang Lab, версия за 22 августа 2019 года. Файл с расширением .exe для запуска алгоритма был получен с помощью компилятора g++. В качестве входных данных я брала наборы данных “Full structures v1.1” и “Interfaces v1.1” из раздела “Docking templates” на сайте Dockground.

* 1. **Таргет-белок**

В качестве таргет-белка (рисунок 3.1) я взяла 1a220A0B\_1 (выделен зеленым цветом) и 1a220A0B\_2 (выделен голубым цветом). Согласно базе данных PDB, белок 1a22 является гормоном роста человека (цепь A), связанным с одиночным рецептором (цепь B).

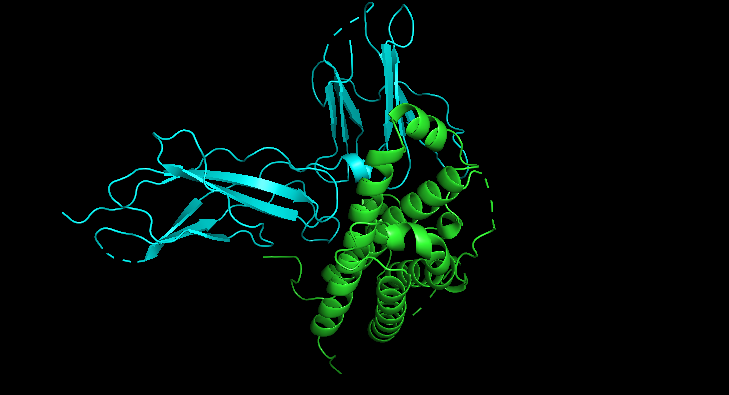


Рисунок 3.1 – Две части таргет-белка 1a22

Запуск алгоритма для таргет-белка я автоматизировала с помощью программы на языке Python. Пример команды:

>TMalign chain\_1.pdb chain\_2.pdb -m matrix.txt,

где TM-align – исполняемый файл, chain\_1.pdb – таргет-белок, chain\_2.pdb – шаблон, matrix.txt – файл, где будет храниться матрица вращения для chain\_1.pdb.

* 1. **Оценка шаблонного моделирования**

Алгоритм TM-align возвращает отчет (рисунок 3.2), где указаны две оценки шаблонного моделирования в зависимости от нормализации. В случае набора данных “Full strucutres v1.1” я брала оценку шаблонного моделирования, нормализованную на длину первого белка – таргет-белка, в случае набора данных “Interfaces v1.1” – нормализованную на длину шаблона. После выравнивания двух частей таргет-белка на две части шаблона, я брала минимальную из двух полученных оценок шаблонного моделирования. Так я получала минимальную оценку выравнивания таргет-белка на шаблон. Также в процессе работы алгоритма создавались для каждого выравнивания файлы с матрицами вращения (рисунок 3.3).

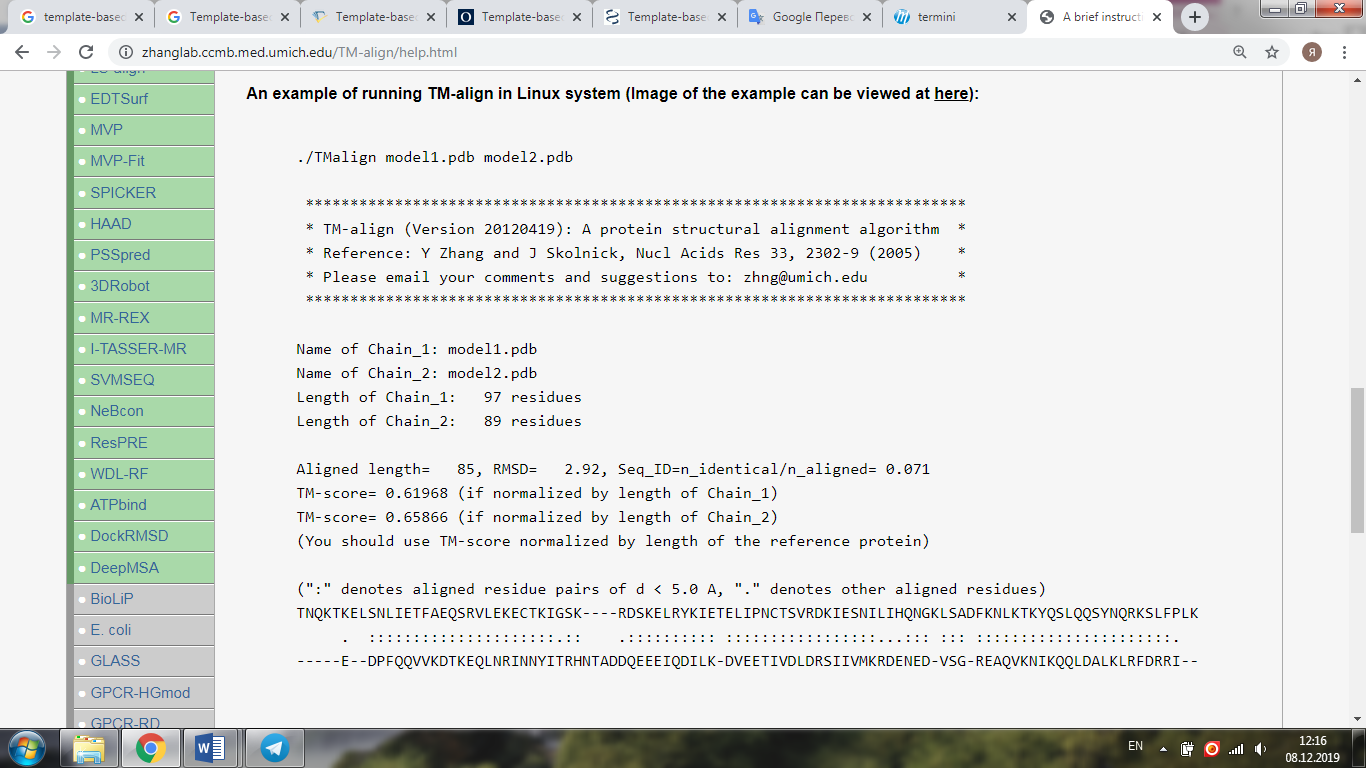


Рисунок 3.2 – Пример отчета, выдаваемый алгоритмом TM-align

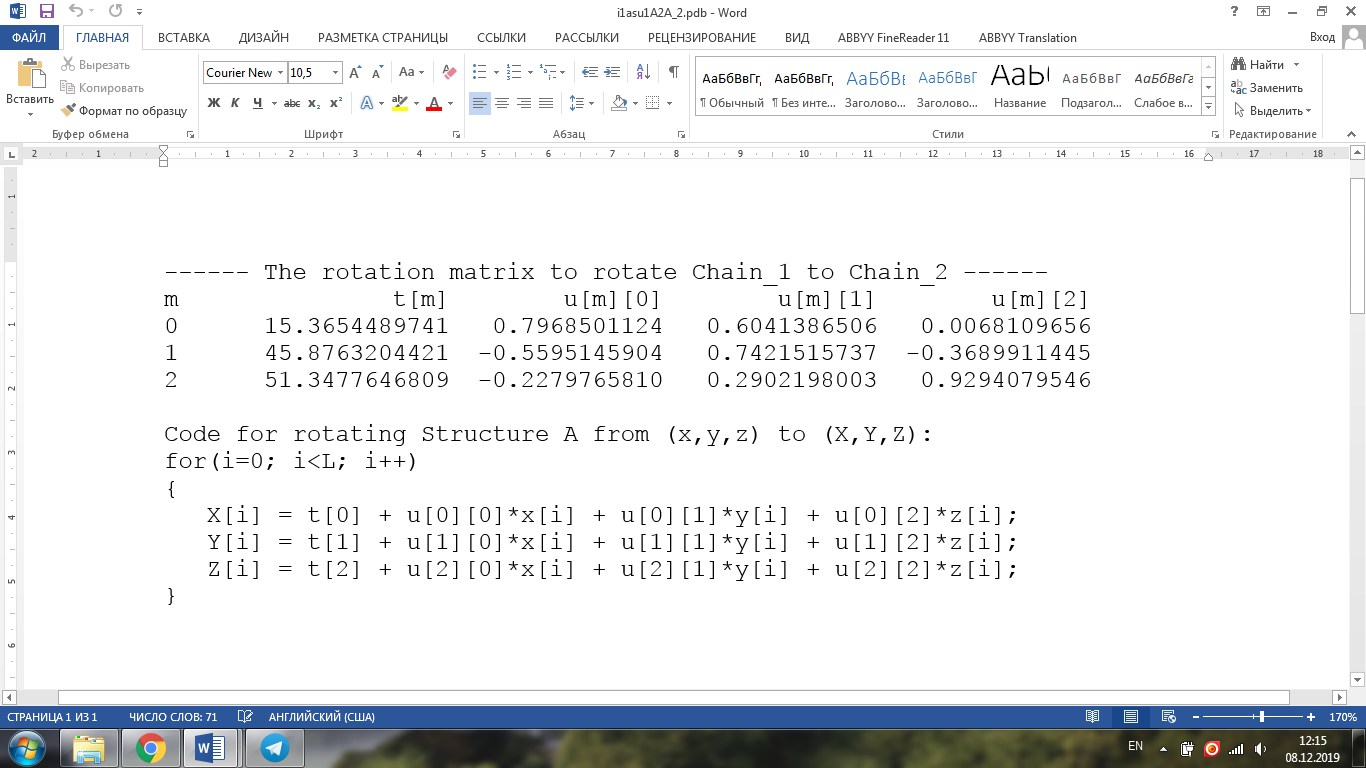


Рисунок 3.3 – Пример матрицы вращения, выдаваемой алгоритмом TM-align

* 1. **Топ-10**

На основании запуска алгоритма для определенного выше таргет-белка и библиотеки шаблонов я составила топ-10 лучших (таблица 3.1, таблица 3.3) и топ-10 худших (таблица 3.2, таблица 3.4) выравниваний на основе полученных оценок шаблонного моделирования.

Как оказалось, в библиотеке шаблонов находился и сам таргет-белок. Для него оценка шаблонного моделирования равна 1.0, то есть алгоритм для одного и того же белка показывает правильную оценку шаблонного моделирования и единичную матрицу вращения.

Таблица 3.1 – Топ-10 лучших выравниваний на шаблоны, являющиеся полными структурами

|  |  |  |  |
| --- | --- | --- | --- |
| Место | Шаблон-1 | Шаблон-2 | TM-score |
| 1 | 1a220A0B\_1 | 1a220A0B\_2 | 1.0 |
| 2 | 1bp30A0B\_1 | 1bp30A0B\_2 | 0.845 |
| 3 | 3n061A1B\_1 | 3n061A1B\_2 | 0.80617 |
| 4 | 1pvh0A0B\_2 | 1pvh0A0B\_1 | 0.6122 |
| 5 | 3d871A1B\_1 | 3d871A1B\_2 | 0.59429 |
| 6 | 3se41A1B\_2 | 3se41A1B\_1 | 0.57576 |
| 7 | 3se41B1C\_1 | 3se41B1C\_2 | 0.57576 |
| 8 | 3d851C1D\_1 | 3d851C1D\_2 | 0.55323 |
| 9 | 3og41A1B\_1 | 3og41A1B\_2 | 0.51067 |
| 10 | 1iar0A0B\_1 | 1iar0A0B\_2 | 0.50857 |

Таблица 3.2 – Топ-10 худших выравниваний на шаблоны, являющиеся полными структурами

|  |  |  |  |
| --- | --- | --- | --- |
| Место | Шаблон-1 | Шаблон-2 | TM-score |
| 1 | 1dfn0A0B\_2 | 1dfn0A0B\_1 | 0.07515 |
| 2 | 2f910A0B\_2 | 2f910A0B\_1 | 0.08381 |
| 3 | 1clv0A0I\_2 | 1clv0A0I\_1 | 0.0854 |
| 4 | 1clv0A0I\_1 | 1clv0A0I\_2 | 0.08584 |
| 5 | 1dxg0A0B\_1 | 1dxg0A0B\_2 | 0.08673 |
| 6 | 1dxg0A0B\_2 | 1dxg0A0B\_1 | 0.08778 |
| 7 | 2ds80A0B\_1 | 2ds80A0B\_2 | 0.08808 |
| 8 | 1dfn0A0B\_1 | 1dfn0A0B\_2 | 0.08815 |
| 9 | 3f1f1G14\_2 | 3f1f1G14\_1 | 0.08867 |
| 10 | 2xtt1A1B\_1 | 2xtt1A1B\_2 | 0.08924 |

Таблица 3.3 – Топ-10 лучших выравниваний на шаблоны, являющиеся интерфейсами

|  |  |  |  |
| --- | --- | --- | --- |
| Место | Шаблон-1 | Шаблон-2 | TM-score |
| 1 | i1a220A0B\_1 | i1a220A0B\_2 | 1.0 |
| 2 | i3n061A1B\_1 | i3n061A1B\_2 | 0.85223 |
| 3 | i1bp30A0B\_1 | i1bp30A0B\_2 | 0.81014 |
| 4 | i3se41B1C\_1 | i3se41B1C\_2 | 0.67145 |
| 5 | i3d871A1B\_1 | i3d871A1B\_2 | 0.65577 |
| 6 | i3d871C1D\_1 | i3d871C1D\_2 | 0.60893 |
| 7 | i1pvh0A0B\_2 | i1pvh0A0B\_1 | 0.59036 |
| 8 | i3d851C1D\_1 | i3d851C1D\_2 | 0.56639 |
| 9 | i3g9v1A1B\_2 | i3g9v1A1B\_1 | 0.56088 |
| 10 | i3s9d1A1B\_1 | i3s9d1A1B\_2 | 0.53908 |

Таблица 3.4 – Топ-10 худших выравниваний на шаблоны, являющиеся интерфейсами

|  |  |  |  |
| --- | --- | --- | --- |
| Место | Шаблон-1 | Шаблон-2 | TM-score |
| 1 | i1sx40A0O\_2 | i1sx40A0O\_1 | 0.10335 |
| 2 | i1u6g0A0C\_1 | i1u6g0A0C\_2 | 0.11276 |
| 3 | i1we30F0T\_2 | i1we30F0T\_1 | 0.11552 |
| 4 | i1u6g0A0C\_2 | i1u6g0A0C\_1 | 0.11753 |
| 5 | i1wf40c0q\_2 | i1wf40c0q\_1 | 0.11946 |
| 6 | i2ix30A0B\_1 | i2ix30A0B\_2 | 0.12436 |
| 7 | i2ch70A0B\_2 | i2ch70A0B\_1 | 0.12499 |
| 8 | i3k6s1A1B\_1 | i3k6s1A1B\_2 | 0.12611 |
| 9 | i1o940A0B\_2 | i1o940A0B\_1 | 0.12617 |
| 10 | i2ch70A0B\_1 | i2ch70A0B\_2 | 0.12885 |

* 1. **Визуализация молекул**

Я визуализировала лучшее и худшее выравнивания с помощью соответствующих матриц вращения, чтобы посмотреть результаты выравнивания алгоритма TM-align. Действительно, оценка шаблонного моделирования меньше 0.2 соответствует случайно выбранным неродственным белкам.

В программу pyMOL для визуализации выравнивания надо загрузить таргет-белок и шаблон, а затем применить команду cmd.transform\_object к таргет-белку. Пример команды:

cmd.transform\_object("1a220A0B\_1",[-0.3122693437, -0.8478175770, -0.4285944623, 120.0265438426, 0.3524983502, -0.5223542466, 0.7764605297, -67.3976053379, -0.8821750224, 0.0913859791, 0.4619695150, -14.2188984012, 0,0,0,1],homogenous=0).

***“Full structures v1.1”***

Далее представлены визуализации выравнивания таргет-белка 1a22 на шаблоны 1bp3 (лучшее выравнивание) (рисунок 3.4, рисунок 3.5, рисунок 3.6) и 1dfn (худшее выравнивание) (рисунок 3.7, рисунок 3.8, рисунок 3.9), являющиеся полными структурами.



Рисунок 3.4 – Выровненные 1bp30A0B\_1 (голубой) и 1a220A0B\_1 (зеленый)



Рисунок 3.5 – Выровненные 1bp30A0B\_2 (фиолетовый) и 1a220A0B\_2 (желтый)



Рисунок 3.6 – Выровненные 1bp30A0B (фиолетовый) и 1a220A0B (желтый)



Рисунок 3.7 – Выровненные 1dfn0A0B\_2 (голубой) и 1a220A0B\_1 (зеленый)

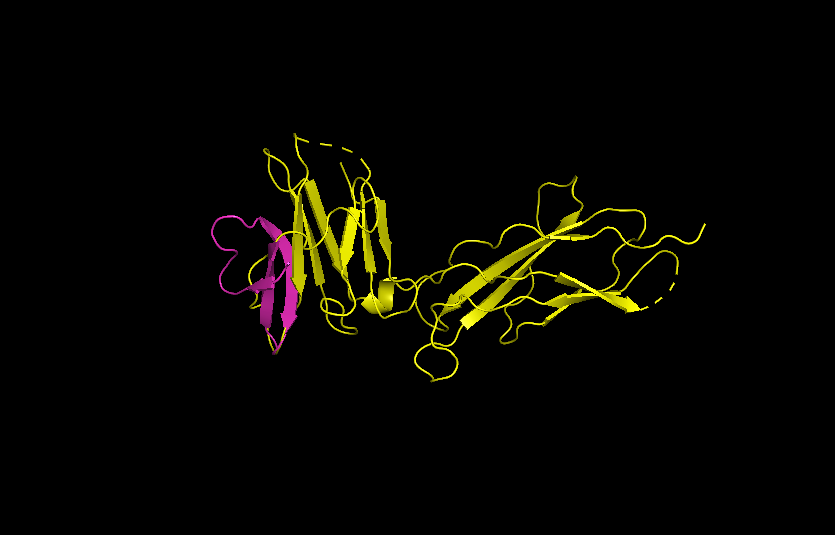


Рисунок 3.8 – Выровненные 1dfn0A0B\_1 (фиолетовый) и 1a220A0B\_2 (желтый)

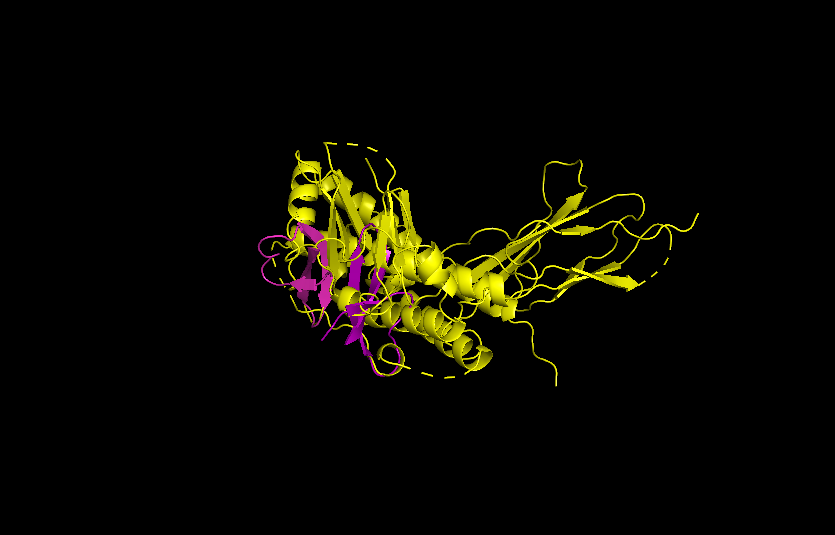


Рисунок 3.9 – Выровненные 1dfn0A0B (фиолетовый) и 1a220A0B (желтый)

***“Interfaces v1.1”***

Далее представлены визуализации выравнивания таргет-белка 1a22 на шаблоны i3n06 (лучшее выравнивание) (рисунок 3.10, рисунок 3.11, рисунок 3.12) и i1sx4 (худшее выравнивание) (рисунок 3.13, рисунок 3.14, рисунок 3.15), являющиеся интерфейсами.

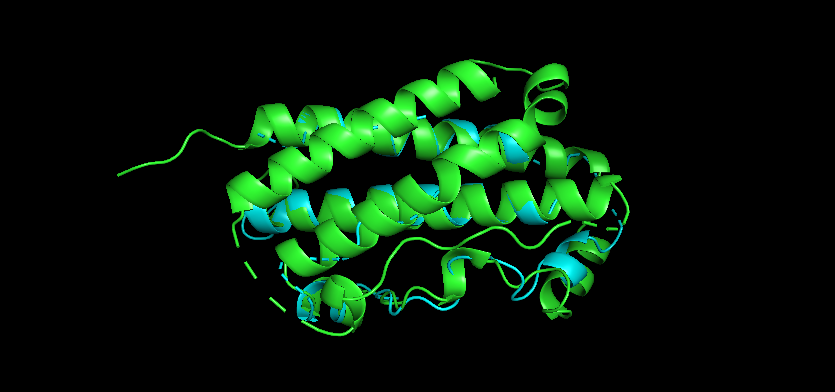


Рисунок 3.10 – Выровненные i3n061A1B\_1 (голубой) и 1a220A0B\_1 (зеленый)

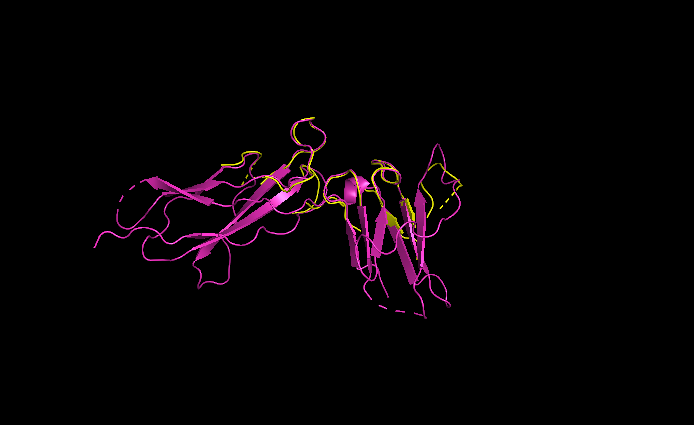


Рисунок 3.11 – Выровненные i3n061A1B\_2 (желтый) и 1a220A0B\_2 (фиолетовый)

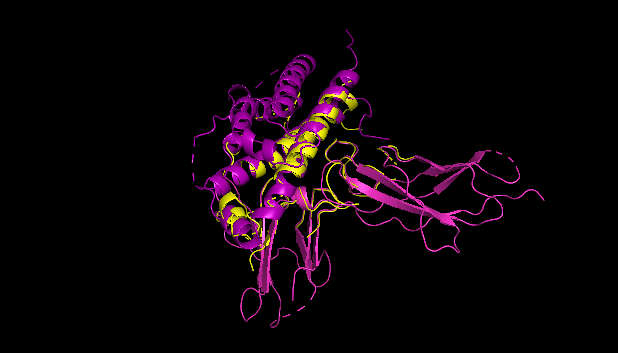


Рисунок 3.12 – Выровненные i3n061A1B (желтый) и 1a220A0B (фиолетовый)



Рисунок 3.13 - Выровненные i1sx40A0O\_2 (голубой) и 1a220A0B\_1 (зеленый)



Рисунок 3.14 - Выровненные i1sx40A0O\_1 (желтый) и 1a220A0B\_2 (фиолетовый)



Рисунок 3.15 - Выровненные i1sx40A0O(желтый) и 1a220A0B (фиолетовый)

Действительно, алгоритм TM-align распознает идентичные структуры (оценка шаблонного моделирования равна 1.0), схожие структуры (оценка шаблонного моделирования больше 0.5) и случайные неродственные структуры (оценка шаблонного моделирования меньше 0.2), а программа pyMOL наглядно это показывает.

* 1. **Среднеквадратичное отклонение**

Для вычисления среднеквадратичного отклонения структурного выравнивания я сделала следующее с помощью программы pyMOL:

1. применила матрицы вращения для двух частей таргет-белков,
2. сохранила две части таргет-белка как один белок и две части шаблона как один белок,
3. помощью команды rms\_cur вычислила среднеквадратичное отклонение.

Далее сравниваются оценка шаблонного моделирования и среднеквадратичное отклонение для набора “Full structures v1.1” (таблица 3.5), для набора “Interfaces v1.1” (таблица 3.6).

Таблица 3.5 – Сравнение оценок для набора “Full structures v1.1”

|  |  |
| --- | --- |
| TM-score | RMSD |
| 1.0 | 0.0 |
| 0.845 | 6.187 |
| 0.07515 | 27.978 |

Таблица 3.6 – Сравнение оценок для набора “Interfaces v1.1”

|  |  |
| --- | --- |
| TM-score | RMSD |
| 1.0 | 0.0 |
| 0.85223 | 18.177 |
| 0.10335 | 130.143 |

Как видно из таблицы 6, в случае интерфейсов среднеквадратичное отклонение может быть высоким из-за локальных структурных отклонений. Данное сравнение подтверждает теорию из пункта 2.2.

**ГЛАВА 4. УСОВЕРШЕНСТВОВАНИЕ АЛГОРИТМА**

**4.1 Обработка входных данных**

В наборах данных полных структур и интерфейсов имеется папка ‘info’ с информацией о результатах алгоритма выравнивания TM-align всех таргет-белков на всю библиотеку шаблонов, причем отобраны только те выравнивания, оценка шаблонного моделирования которых больше 0.4.

Чтобы удостовериться в имеющейся в папке ‘info’ информации, я запустила алгоритм для этих отобранных пар. Как выяснилось, в этой папке за конечную оценку шаблонного моделирования для пары таргет-белок-шаблон бралась максимальная из двух оценок. Я решила парной оценкой шаблонного моделирования считать минимальную из двух оценок, полученных во время выравнивания двух частей таргет-белка на шаблон. Это позволило ужесточить алгоритм.

**4.2 Текущий алгоритм**

Изначально алгоритм для предсказания структуры белкового комплекса на основе белок-белковых интерфейсов в качестве структурных шаблонов состоит из следующих шагов:

1. Для двух частей таргет-белка использую алгоритм TM-align. Выравниванию две части таргет-белка на все имеющиеся шаблоны. В качестве оценки шаблонного моделирования для одного выравнивания беру оценку, нормализованную на длину таргет-белка в случае, когда шаблонами являются полные структуры, и нормализованную на длину шаблона в случае, когда шаблонами являются интерфейсы. В качестве оценки шаблонного моделирования для таргет-белка беру минимальную оценку из двух оценок выравнивания двух частей таргет-белка.
2. Полученные выравнивания сортирую по убыванию оценки шаблонного моделирования.
3. Применяю для таргет-белка матрицы трансформации.
4. Считаю полное среднеквадратичное отклонение атомных положений и среднеквадратичное отклонение атомных положений между лигандами шаблона и трансформированного таргет-белка.
5. Модифицирую список выравниваний: оставляю только те выравнивания, среднеквадратичное отклонение атомных положений которых меньше или равно 10.

На выходе – топ лучших выравниваний для таргет-белка.

**4.3 Улучшение алгоритма**

Для улучшения имеющегося алгоритма я решила модифицировать библиотеку шаблонов следующим образом: сформировала набор данных - пересечение шаблонов, то есть такие файлы белков с расширением .pdb, для которых есть и полная структура, и интерфейс. За конечную оценку шаблонного моделирования я брала минимальную из двух – оценка выравнивания таргет-белка на шаблон-интерфейс и оценка выравнивания таргет-белка на шаблон-полную структуру.

**4.4 Типы графиков**

Для оценки точности алгоритма я построила два типа графиков. Первый – точечный график соответствия среднеквадратичного отклонения атомных положений и оценки шаблонного моделирования. Второй – столбчатая диаграмма коэффициентов успешных выравниваний.

*Точечный график*

Для построения данного графика я беру пары ‘среднеквадратичное отклонение атомных положений - оценка шаблонного моделирования’ по всем таргет-белкам и строю точечный график, чтобы посмотреть, как в моем алгоритме среднеквадратичное отклонение атомных положений соответствует оценке шаблонного моделирования.

*Столбчатая диаграмма*

Для каждого таргет-белка у меня есть список всех шаблонов, выравнивания на которые дали оценку шаблонного моделирования больше или равную 0.4. Эти списки отсортированы по убыванию оценки шаблонного моделирования, то есть от самого лучшего выравнивания к самому худшему.

Я прохожусь по всем таргет-белкам и считаю, у скольких таргет-белков в топ-N (N = {1, 5, 10, 50, 100, n}, где n – количество всех таргет-белков) есть результаты со среднеквадратичным отклонением атомных положений меньше или равным 10.0, и эту сумму делю на общее количество таргет-белков.

**4.5 Графики для изначального алгоритма**

Далее представлены графики для изначального алгоритма без улучшения.

* Точечный график

*Среднеквадратичное отклонение атомных положений между лигандами*

а) Набор данных ‘Full structures’ (рисунок 4.1)

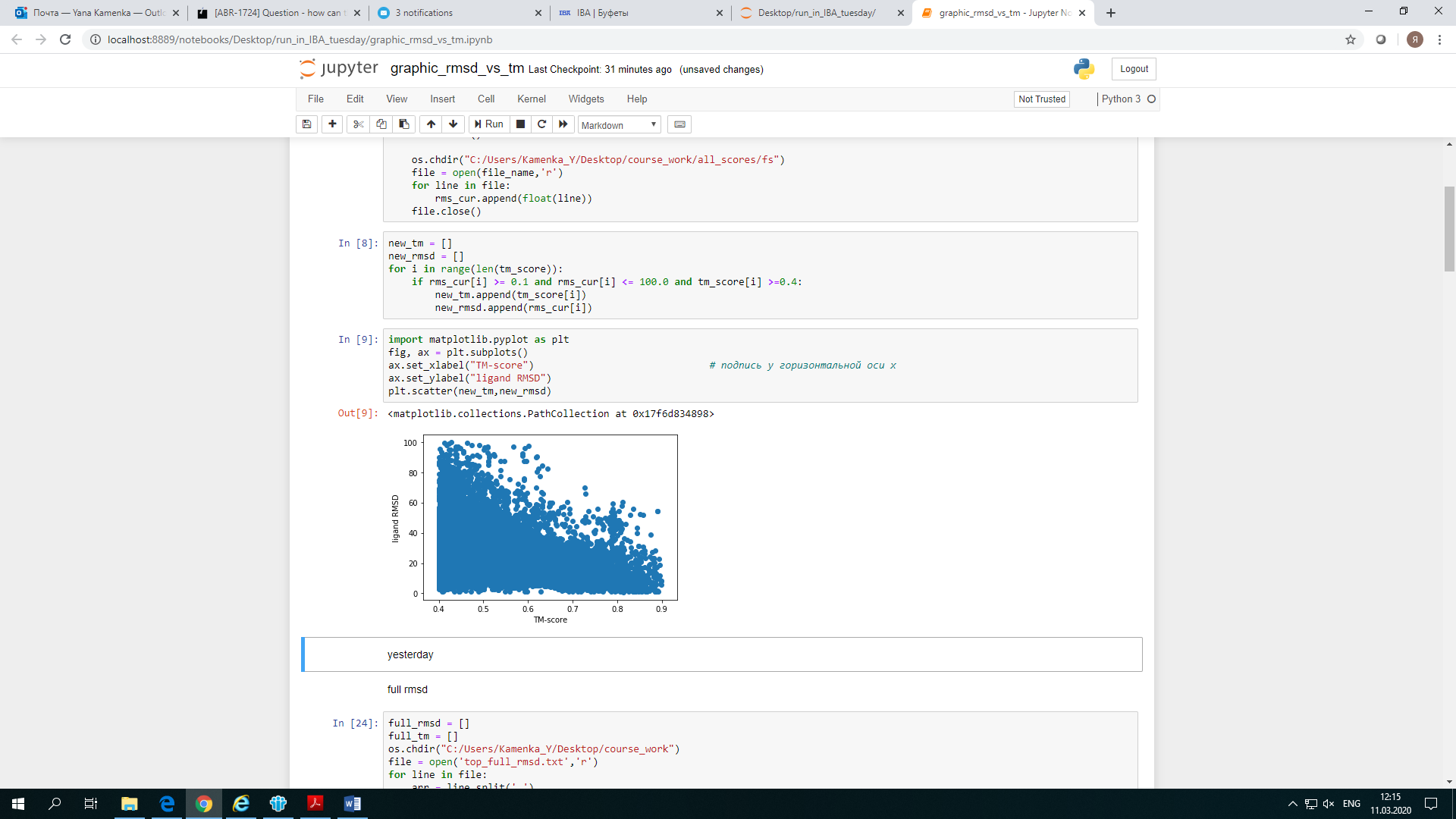


Рисунок 4.1 – Точечный график для «Full structures»

б) Набор данных ‘Interfaces’ (рисунок 4.2)

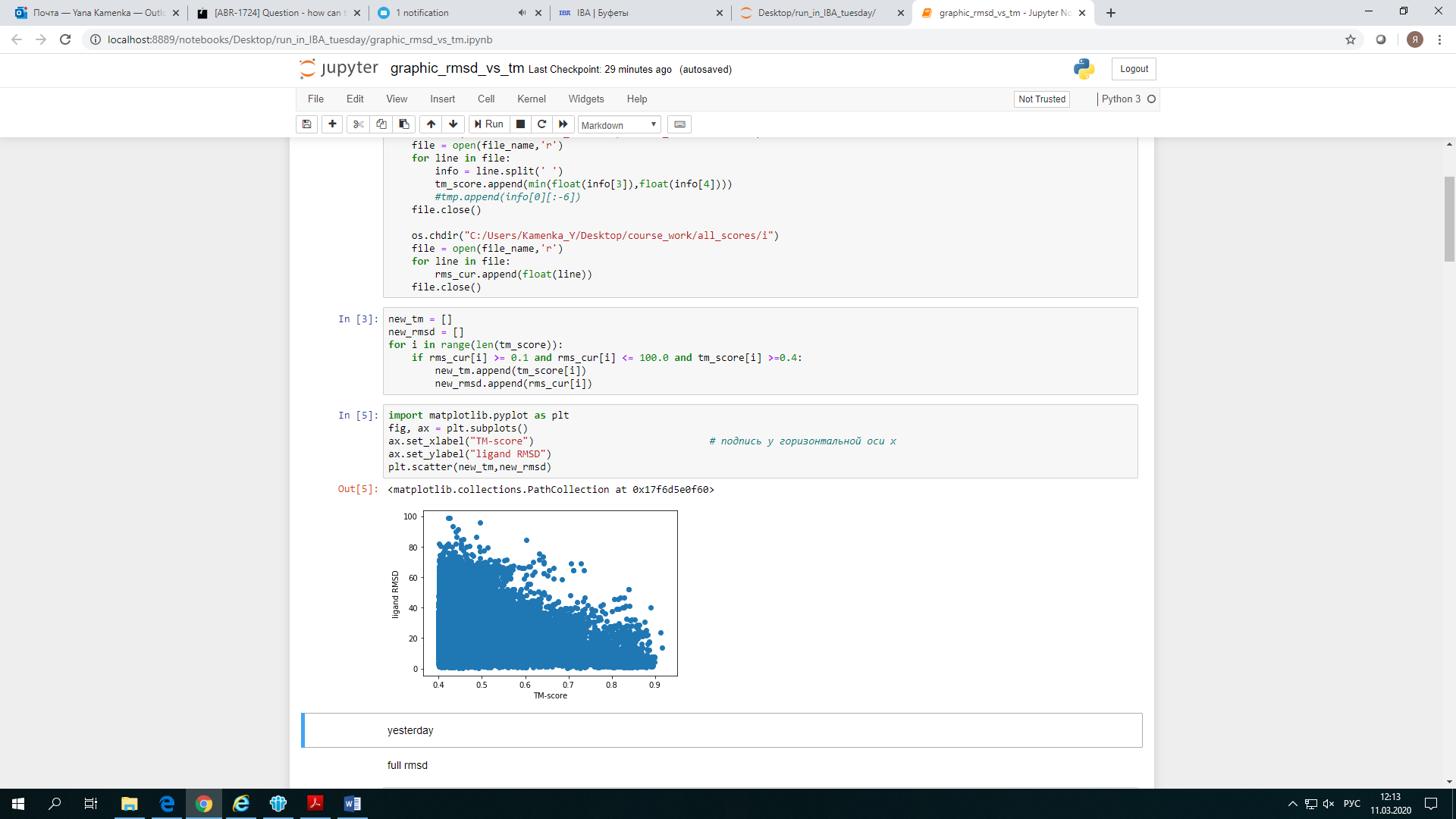


Рисунок 4.2 – Точечный график для «Interfaces»

*Полное среднеквадратичное отклонение атомных положений*

а) Набор данных ‘Full structures’ (рисунок 4.3)

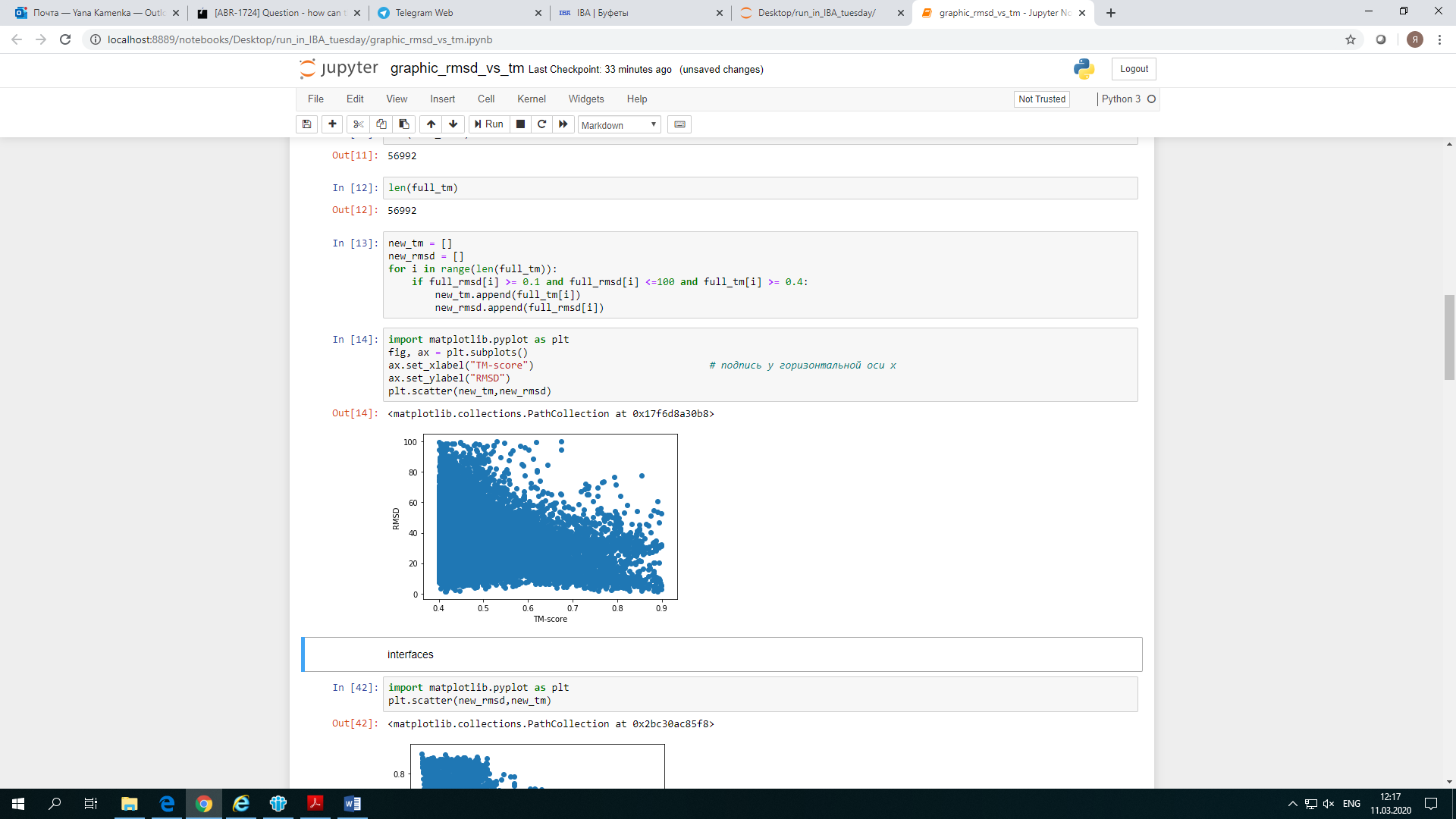


Рисунок 4.3 – Точечный график для «Full structures»

б) Набор данных ‘Interfaces’ (рисунок 4.4)

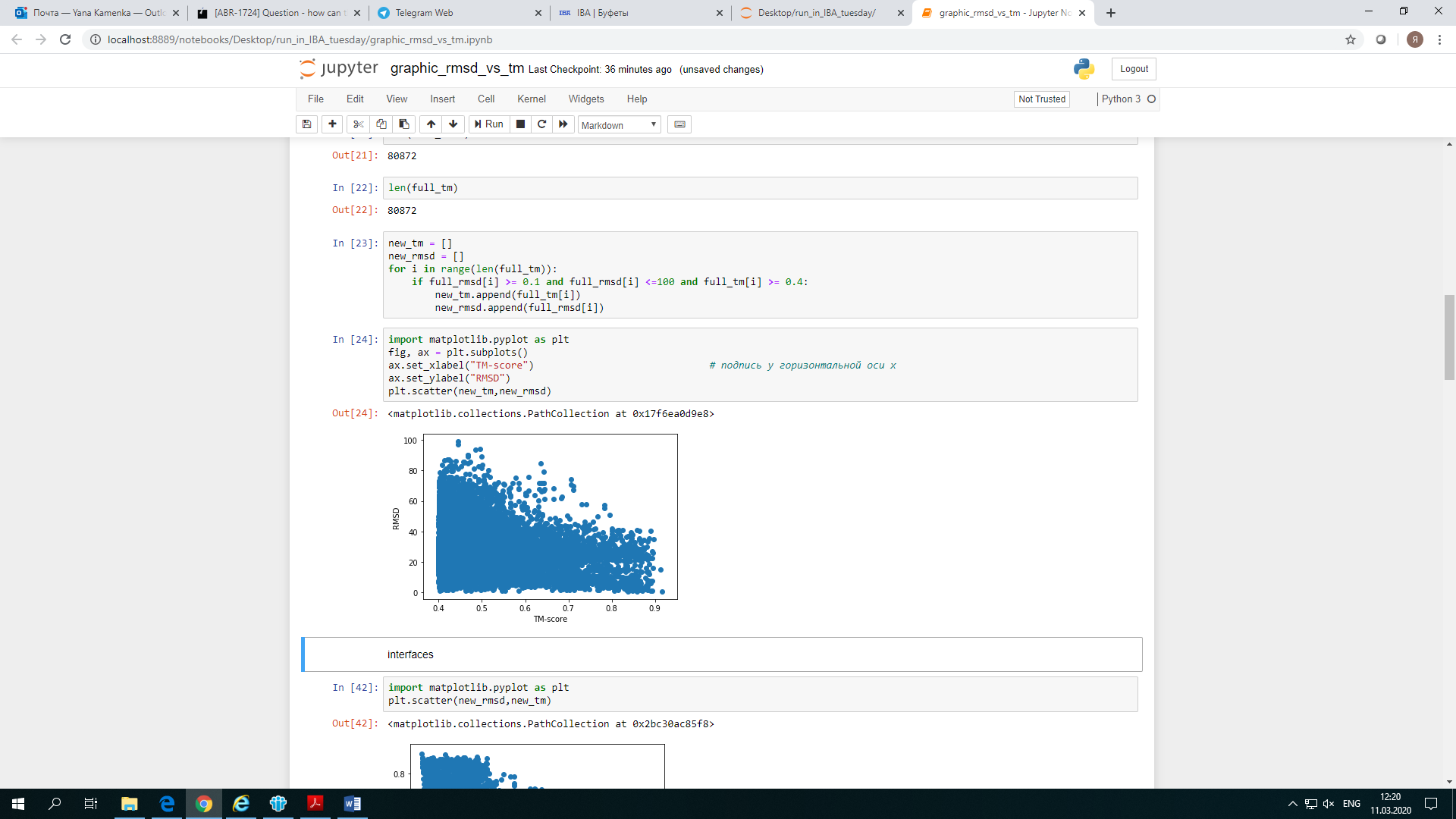


Рисунок 4.4 – Точечный график для «Interfaces»

* Столбчатая диаграмма

*Среднеквадратичное отклонение атомных положений между лигандами*

а) Набор данных ‘Full structures’ (рисунок 4.5)

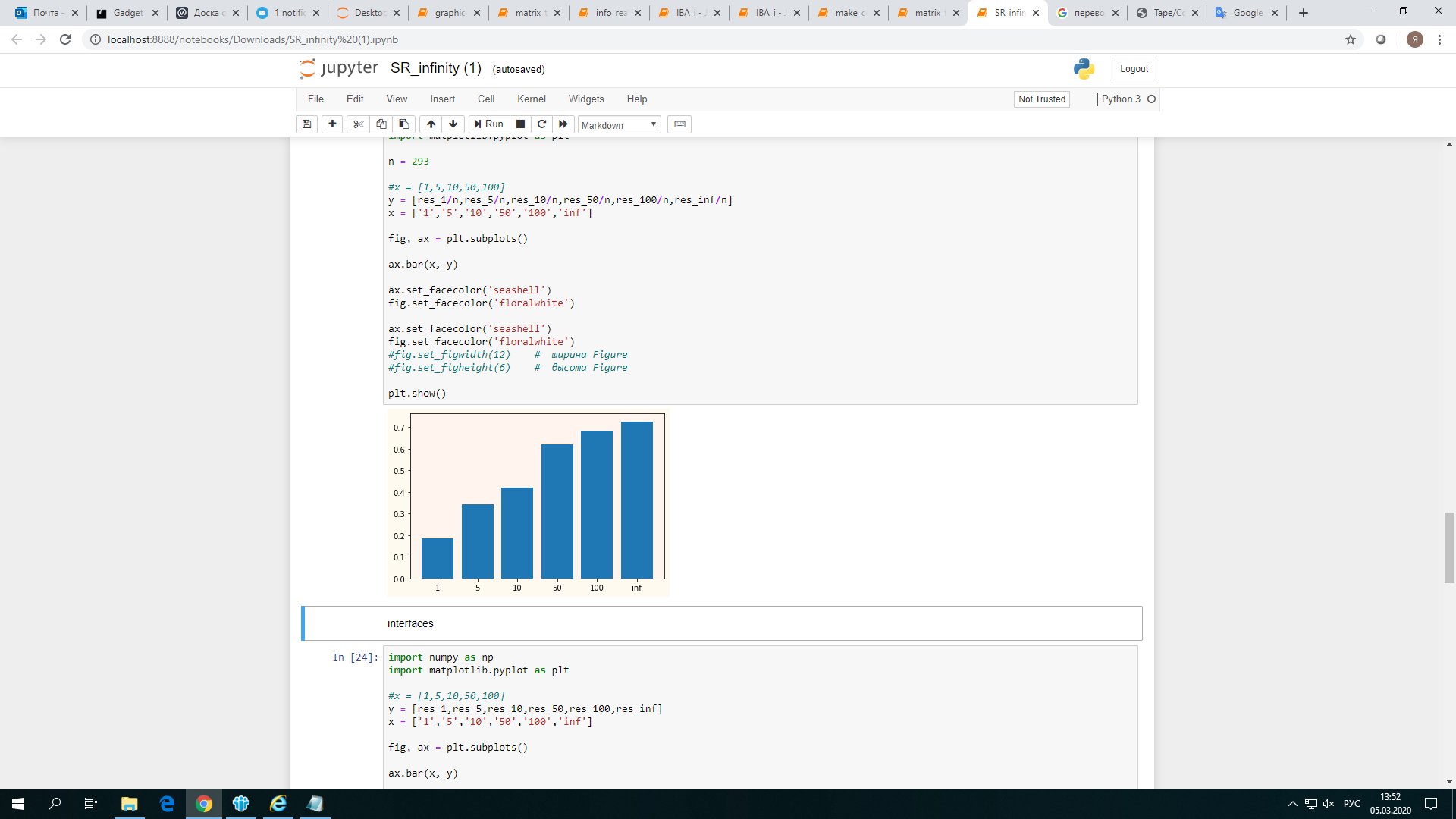


Рисунок 4.5 – Столбчатая диаграмма для «Full structures»

б) Набор данных ‘Interfaces’ (рисунок 4.6)

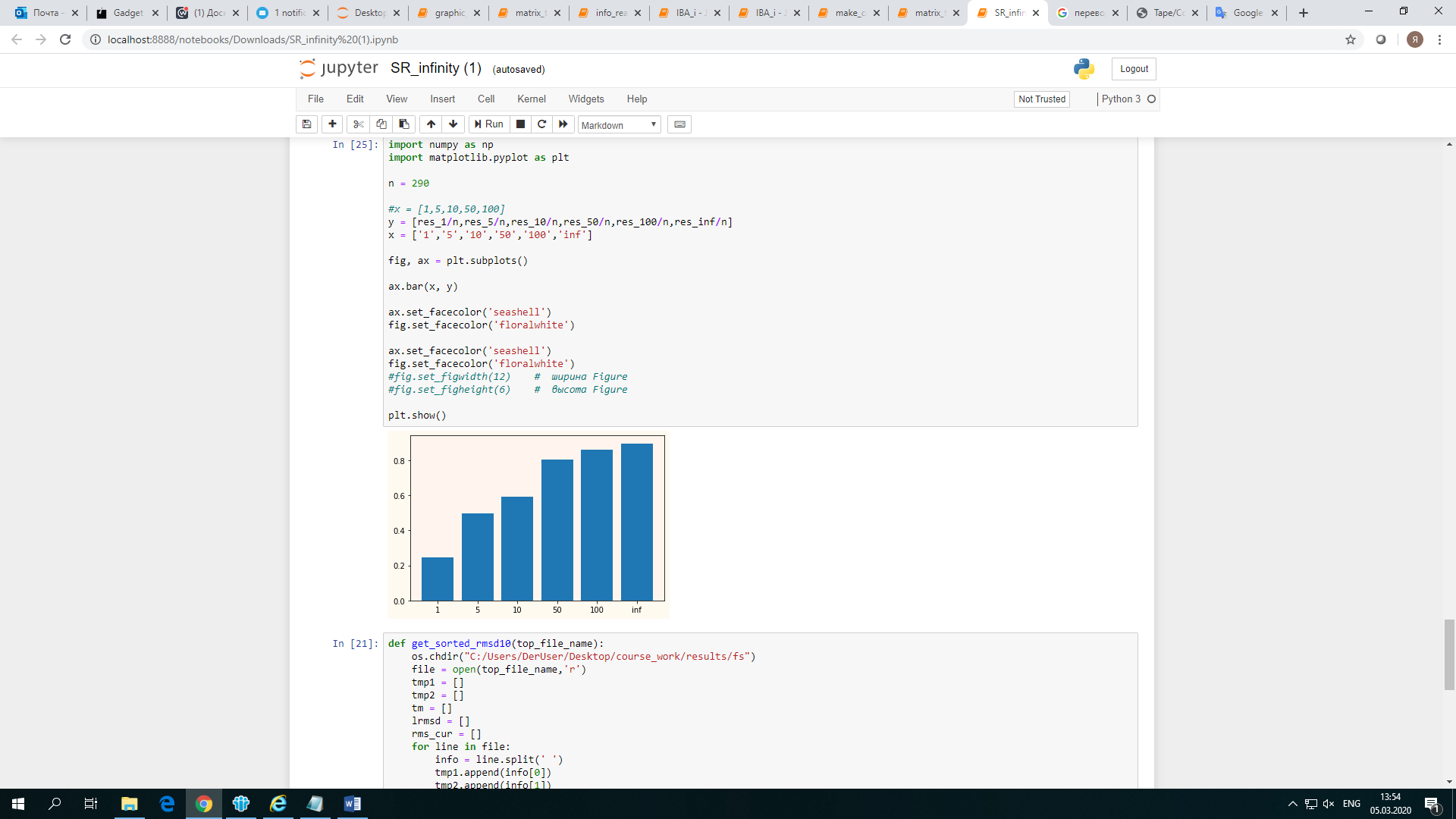


Рисунок 4.6 – Столбчатая диаграмма для «Interfaces»

*Полное среднеквадратичное отклонение атомных положений*

а) Набор данных ‘Full structures’ (рисунок 4.7)

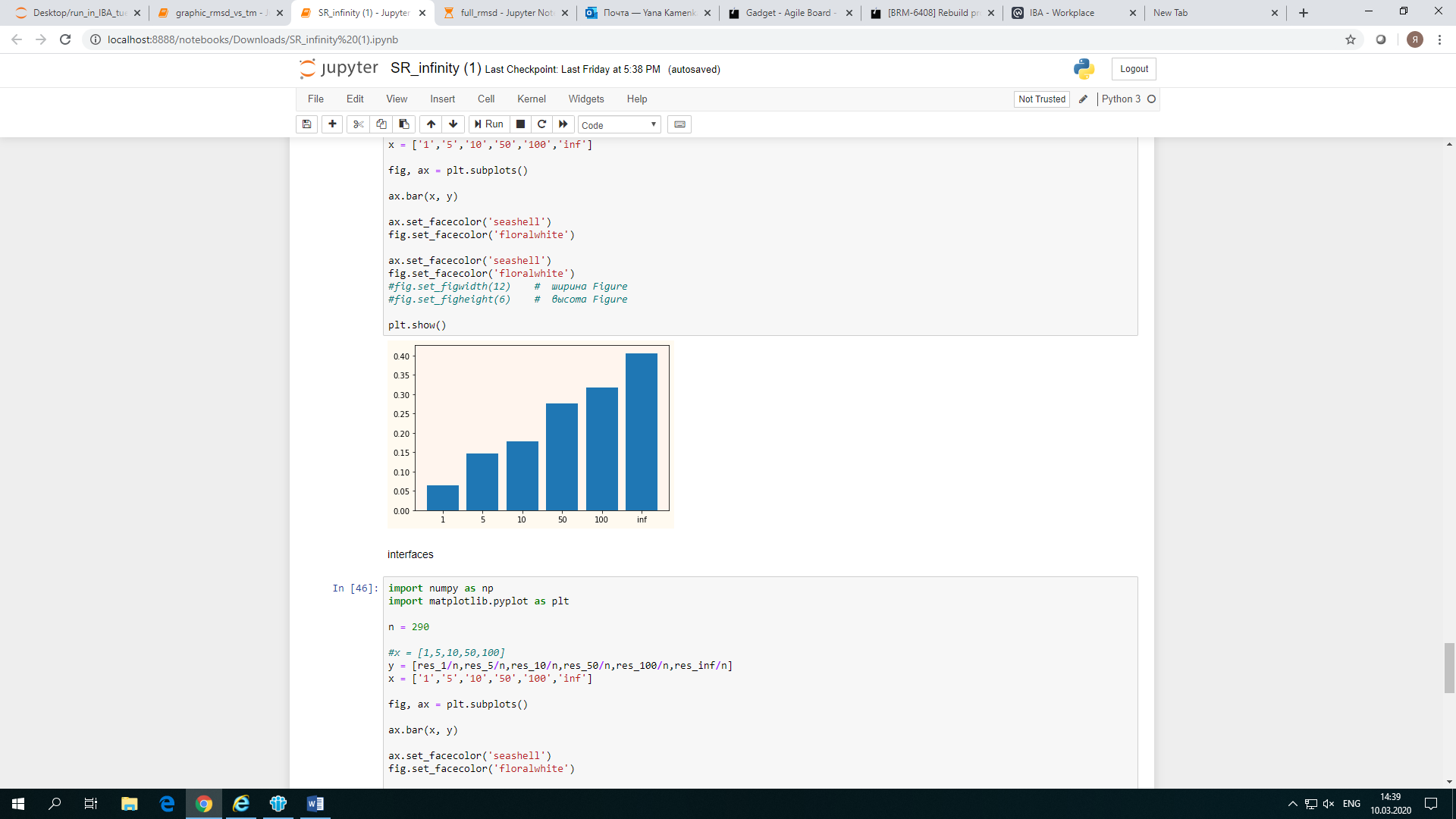


Рисунок 4.7 - Столбчатая диаграмма для «Full structures»

б) Набор данных ‘Interfaces’ (рисунок 4.8)

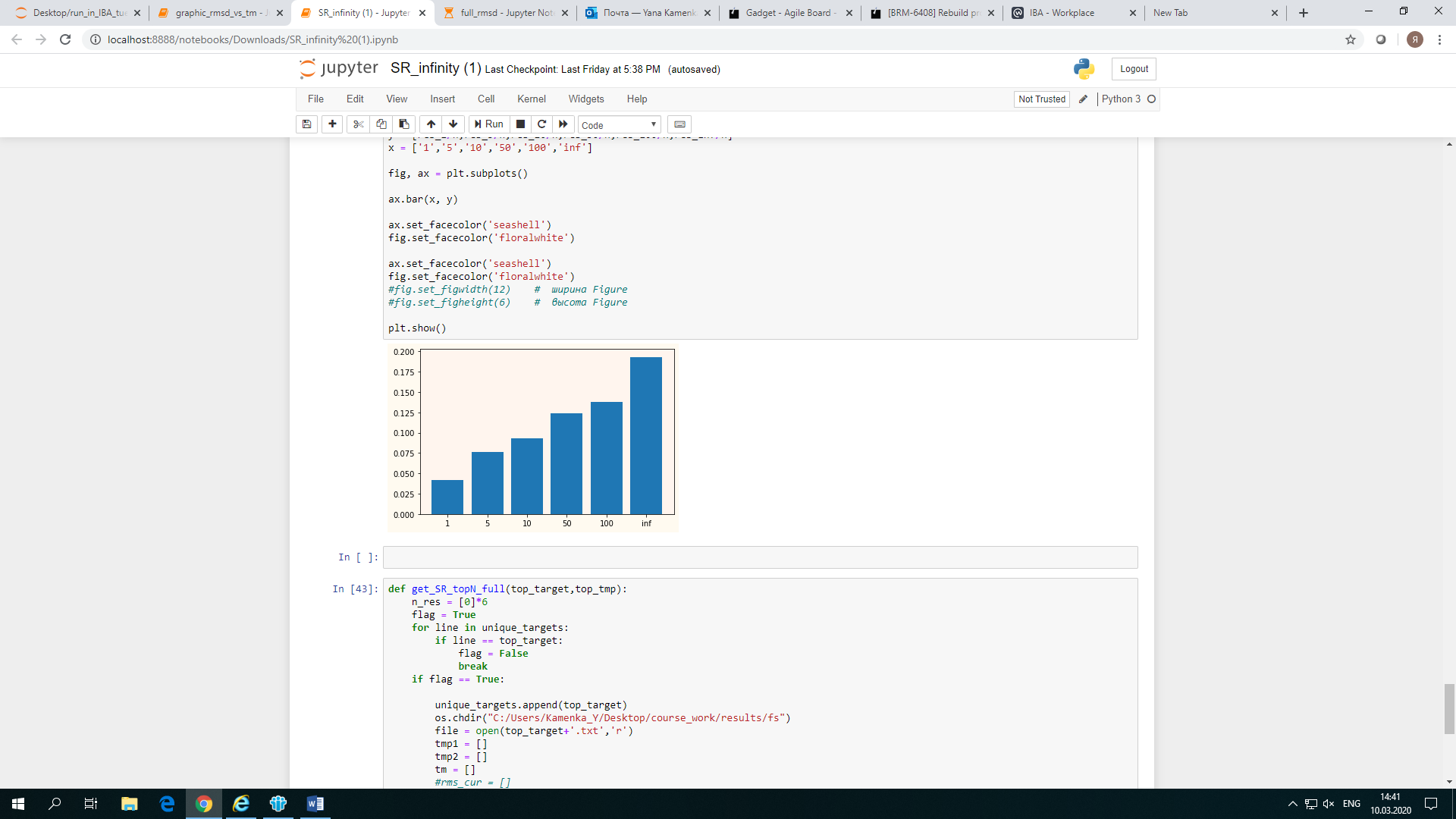


Рисунок 4.8 - Столбчатая диаграмма для «Interfaces»

**4.6 Графики для улучшенного алгоритма**

Далее представлены графики для улучшенного алгоритма.

* Точечный график

*Среднеквадратичное отклонение атомных положений между лигандами (рисунок 4.9)*

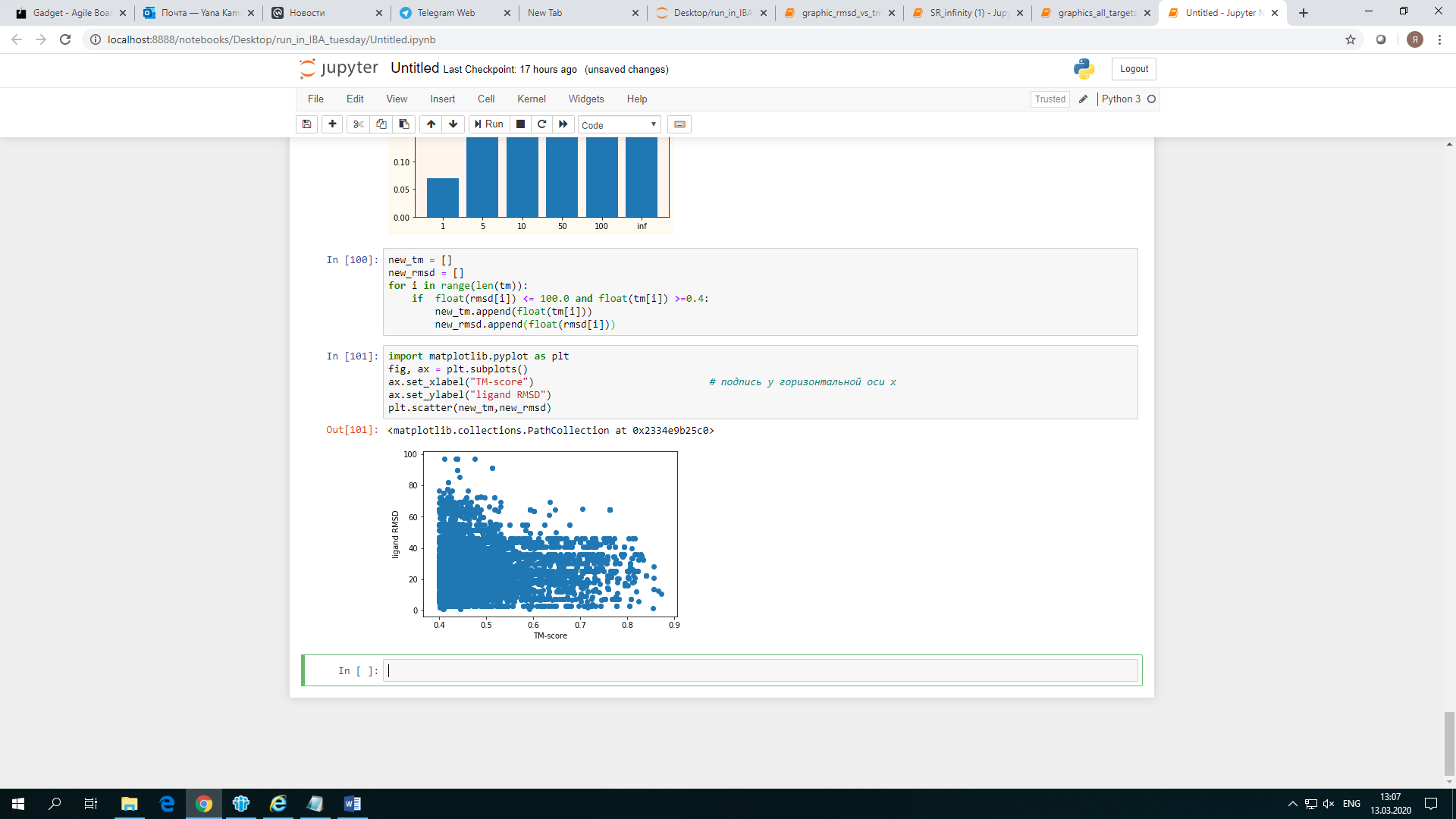


Рисунок 4.9 – Точечный график

*Полное среднеквадратичное отклонение атомных положений (рисунок 4.10)*

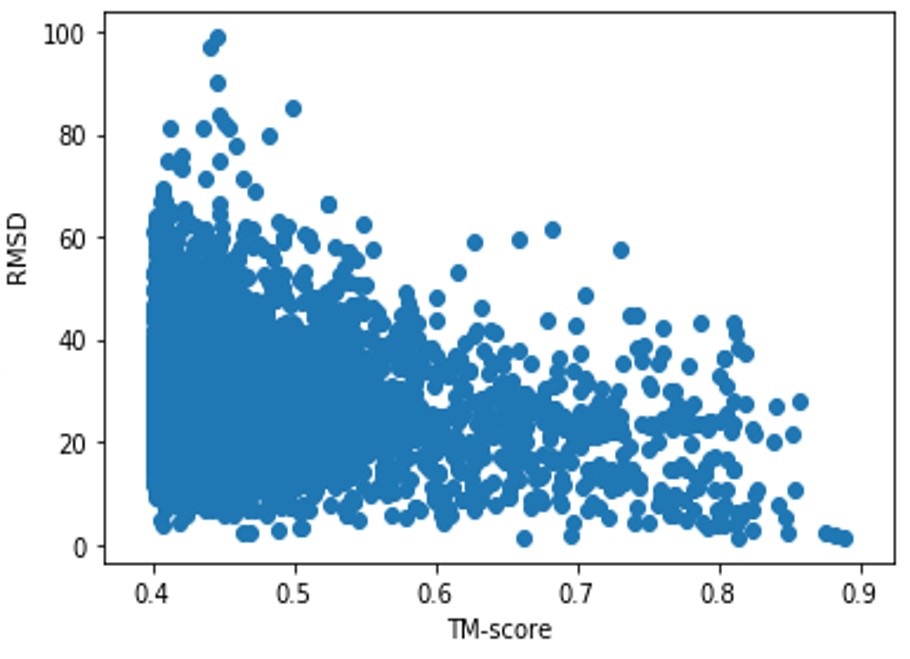


Рисунок 4.10 – Точечный график

* Столбчатая диаграмма

*Среднеквадратичное отклонение атомных положений между лигандами (рисунок 4.11)*

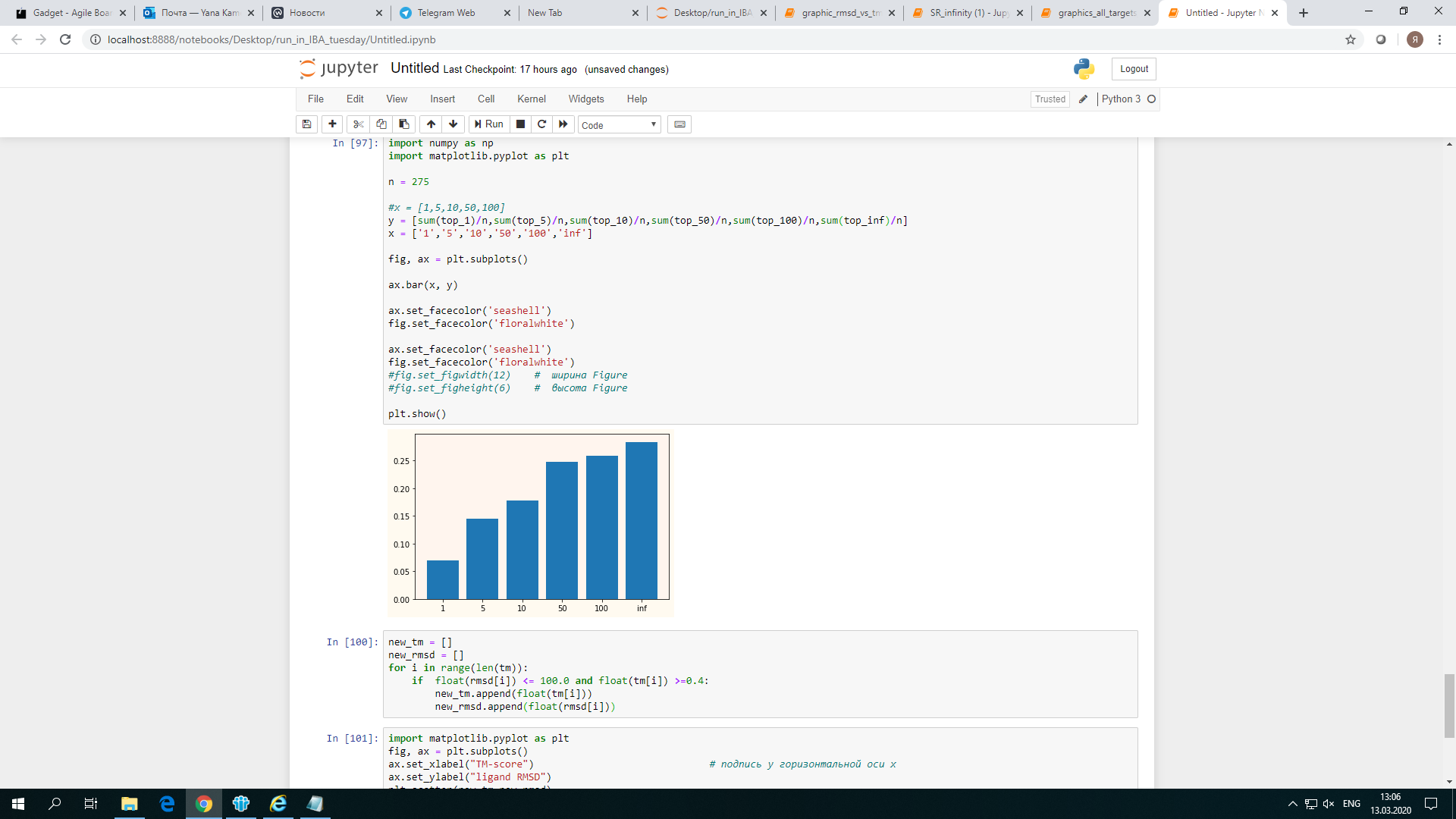


Рисунок 4.11 - Столбчатая диаграмма

*Полное среднеквадратичное отклонение атомных положений (рисунок 4.12)*

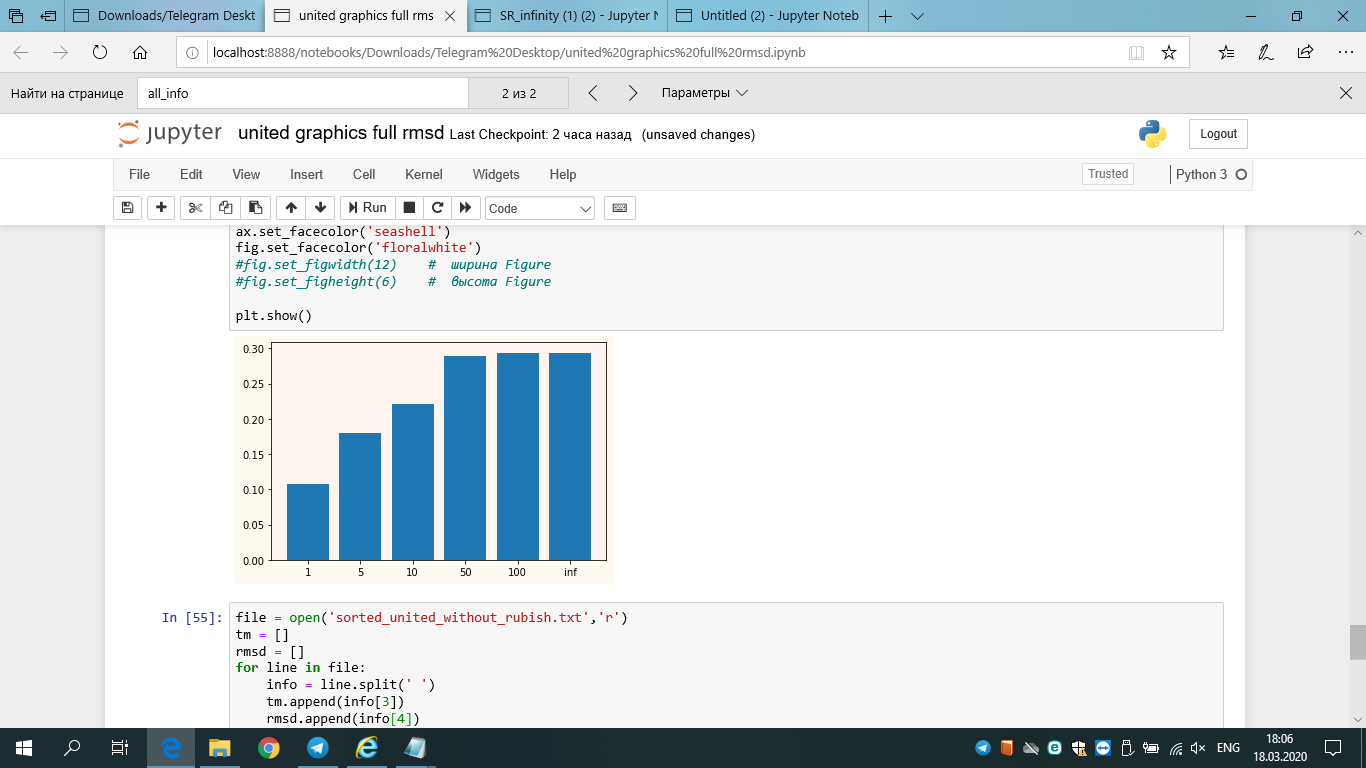


Рисунок 4.12 - Столбчатая диаграмма

**4.7 Сравнение алгоритмов**

Как видно из графиков, столбчатые диаграммы коэффициентов успешных выравниваний показывают лучший результат в случае модифицированного алгоритма. Это показывает, что он лучше находит шаблоны для построения модели за счет выравнивания и на полную структуру, и на интерфейс одного и того же шаблона. Точечные графики соответствия среднеквадратичного отклонения атомных положений и оценки шаблонного моделирования также лучше у модифицированного алгоритма.

**ГЛАВА 5. МОДЕЛИРОВАНИЕ МЕТАБОЛИЧЕСКОГО ПУТИ**

**5.1 Метаболический путь**

Метаболизм представляет из себя совокупность биохимических превращений веществ, поступающих в организм, и взаимопревращения веществ, из которых состоит организм. Метаболическими путями являются цепи последовательных реакций, через которые осуществляются обмен веществ в процессах метаболизма.

В рамках моей дипломной работы под фразой «смоделировать метаболический путь» я подразумеваю «построить белковые комплексы для метаболического пути».

**5.2 База данных UniProt**

UniProt [8] - это открытая база данных последовательностей белков. На данный момент UniProt состоит из четырёх крупных баз данных: база знаний (UniProt KB), архив (UniParc), справочные кластеры (UniRef) и метагеномные данные (UniMes) - и охватывает различные аспекты анализа белковых последовательностей.

В моей дипломной работе используются белки из базы UniProt KB. Она, или если быть точнее, ее секция UniProtKB / Swiss-Prot, содержит обзорные, вручную аннотированные записи. Данная секция содержит всю известную необходимую информации о конкретном белке. Аннотации регулярно проверяются, чтобы соответствовать текущим научным результатам. Требования к составлению аннотации записи включают подробный анализ последовательности белка и данных о нём из научной литературы. Аннотация, как правило, включает следующую информацию: название белка и гена, функция белка, фермент-специфическая информация, такая как каталитическая активность, ко-факторы и каталитические остатки, внутриклеточная локализация, белок-белковые взаимодействия, шаблон экспрессии, местоположение и роль важных доменов и сайтов.

**5.3 Проект KEGG**

KEGG (Kyoto Encyclopedia of Genes and Genomes — Киотская энциклопедия генов и геномов) [2] — веб-ресурс, предоставляющий доступ к ряду биологических баз данных и инструментам для анализа биологических и медицинских данных. Главной задачей веб-ресурса является объединение полученной информации о геномах, биологических и химических процессах, происходящих в живых организмах, знаний о болезнях и существующих лекарствах в единое целое для понимания высокоуровневой организации различных биологических систем, таких как клетка, организм или целая экосистема.

Центром проекта KEGG является KEGG PATHWAY - энциклопедия метаболический путей. Она содержит данные о генных продуктах, связанных в сети белок-белковых взаимодействий, сети ферментов и регуляторов генов. Энциклопедия PATHWAY представлена в виде диаграмм, называемых картами, соответствующих определённым сетям. Существуют карты для клеточных и организменных функций. Данные представлены в виде графических диаграмм, включающих большинство метаболических путей и некоторые из наиболее известных регуляторных путей. В KEGG есть интерактивные схемы клеточных процессов (метаболических реакций, процессов мембранного транспорта, передачи сигналов и др.), метаболические сети и сети белок-белковых взаимодействий.

Среди карт выделяют следующие группы: метаболизм, обработка генетической информации, взаимодействие с окружающей средой, клеточные процессы, организменные системы, заболевания человека, разработка лекарств.

**5.4 Применение алгоритма для метаболического пути**

*Выбор метаболического пути*

Для дипломной работы я искала метаболический путь по следующим соображениям:

1. в нем присутствуют белки;
2. эти белки содержатся в базе данных UniProt Swiss-Prot, то есть являются аннотированными;
3. для этих белков известна их 3D-структура.

Найденным по этим критериям метаболическим путем стал путь K12356 из карты метаболизма. В пути K12356 содержатся три аннотированных белка с известной 3D-структурой – O81498, Q9LVR1, Q94A84.

*Трудности при выборе метаболического пути*

Во время изучения метаболических путей и применения для них моего алгоритма я заметила, что мой алгоритм не для всех путей может найти удачные комплексы. Например, оценка шаблнного моделирования не превышает значения 0.4 или среднеквадратичное отклонение больше значения 10. Я думаю, это связано с тем, что моя библиотека шаблонов может быть недостаточно полной, чтобы находить удачные комплексы для всех белков.

*Удачные комплексы метаболического пути*

Далее представлены удачные комплексы метаболического пути K12356 (рисунок 5.1, рисунок 5.2, рисунок 5.3).

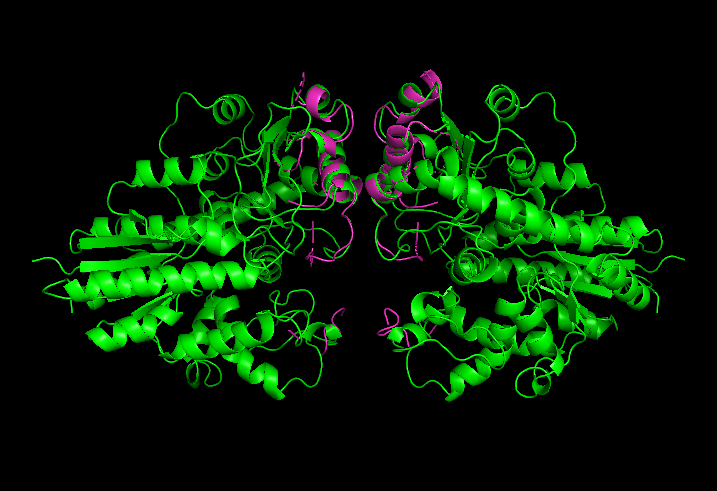


Рисунок 5.1 – Комплекс белков O81498 и Q9LVR1, шаблон 2acv0A0B, TM-score = 0.73982, RMSD = 6.48815345

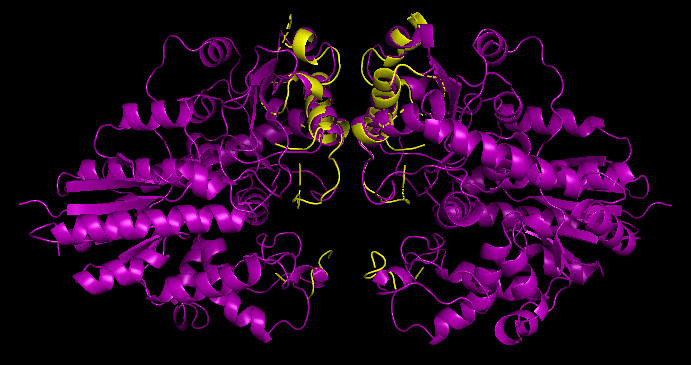


Рисунок 5.2 – Комплекс белков Q94A84 и Q9LVR1, шаблон 2acv0A0B, TM-score = 0.72877, RMSD = 6.7635293

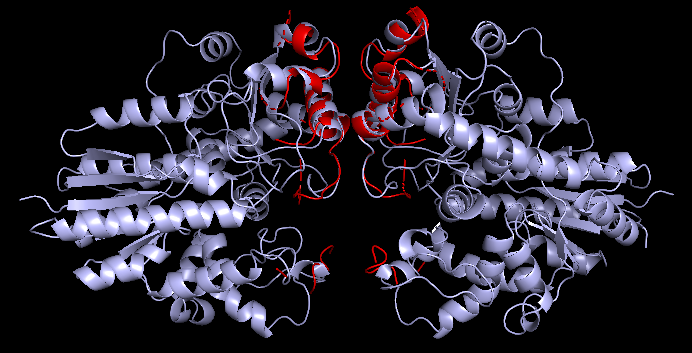


Рисунок 5.3 – Комплекс белков Q81498 и Q94A84, шаблон 2acv0A0B, TM-score = 0.74279, RMSD = 6.48815345

Как видно по визуализации результатов, мой алгоритм успешно находит наилучшие белковые комплексы на основе структурных шаблонов.

**ЗАКЛЮЧЕНИЕ**

В ходе написания дипломной работы я изучила научную литературу по теме докинга, рассмотрела алгоритм структурного выравнивания TM-align, определила входные данные и для этих данных применила алгоритм TM-align, результаты докинга визуализировала с помощью программы pyMOL, доработала алгоритм, оценила его точность, использовала улучшенный алгоритм для моделирования метаболического пути.

Мой алгоритм основан на алгоритме структурного выравнивания TM-align, улучшен благодаря объединению библиотек шаблонов и проверке, что среднеквадратичное отклонение меньше значения 10. Благодаря моему алгоритму появилась возможность моделировать белковые комплексы, например, для метаболических путей, что экономит время и затраты ученых и исследовательских центров на разработку новых лекарственных средств.

В дальнейшем мой алгоритм может быть улучшен благодаря расширению и разнообразию библиотек шаблонов.

**СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ**

1. Dockground [Electronic resource] / Mode of access: http://dockground.compbio.ku.edu/ - Data of access: 15.02.2020
2. KEGG [Electronic resource] / Mode of access: https://www.genome.jp/kegg/ - Data of access: 01.04.2020
3. Template-based protein-protein docking exploiting pairwise interfacial residue restraints / Li C Xue, Joao P.G.L.M. Rodrigues, Drena Dobbs, Vasant Honavar, Alexandre M.J.J. Bonvin // Briefings in Bioinformatics - 2017 – P. 458-466
4. Template-based structure modeling of protein-protein interactions / Andras Szilagyi, Yang Zhang // Curr Opin Struct Biol. – 2015 – P.1-24
5. The PyMOL Molecular Graphics System // Schrödinger, Inc. [Electronic resource] – 2011-2019. – Mode of access: https://pymol.org – Date of access: 15.02.2020.
6. TM-align: a protein structure alignment algorithm based on the TM-score / Yang Zhang, Jeffrey Skolnick // Nucleic Acids Research - 2005 – P. 2302-2309
7. TM-align [Electronic resource] / Mode of access: https://zhanglab.ccmb.med.umich.edu/TM-align/ - Data of access: 15.02.2020
8. UniProt [Electronic resource] / Mode of access: https://www.uniprot.org - Data of access: 06.04.2020