

## Аннотация

Темой данной работы является поиск новых лекарств, возможных побочных эффектов и неспецифичных взаимодействий между биомолекулами. Целью работы являлось построение гибкого программного протокола, способного на основе референсных данных об эффективных и безопасных для человека лекарствах построить поиск наиболее сходного лиганда/мишени/комплекса на основе соответствующих входных данных. В результате такой протокол был создан, проверен и задокументирован. В него также был внедрен пользовательский интерфейс через запуск в командной строке с нужными ключами. Протокол позволяет осуществлять поиск наиболее близких референсных лигандов по их структуре и SMILES идентификатору с помощью молекулярных отпечатков, поиск наиболее схожих референсных белков по их аминокислотным последовательностям и 3D-структурам, а также поиск наиболее похожих комплексов по структуре связывающих карманов. В будущем протокол может дополняться другими способами вычисления подобия, а также на его основе может строиться более сложный поиск, например, с помощью методов машинного обучения.

## Содержание

<b>1</b>	<b>Введение</b>	<b>5</b>
1.1	Обзор литературы . . . . .	6
1.1.1	Полифармакология . . . . .	6
1.1.2	Неспецифичные взаимодействия . . . . .	7
1.1.3	Существующие протоколы . . . . .	8
<b>2</b>	<b>Основное содержание</b>	<b>9</b>
2.1	Общая структура протокола . . . . .	9
2.2	Материалы и методы исследования . . . . .	9
2.2.1	Биологические базы данных . . . . .	9
2.2.2	Коэффициенты Жаккара и Танимото . . . . .	9
2.2.3	Сходство белков по аминокислотной последовательности . . . . .	10
2.2.4	Сходство белков по структуре . . . . .	10
2.2.5	Сходство лигандов по структуре . . . . .	11
2.2.6	Сходство сайтов связывания . . . . .	12
2.3	Описание программного модуля и результаты . . . . .	14
2.3.1	Извлечение данных . . . . .	14
2.3.2	Сходство белков по аминокислотной последовательности . . . . .	15
2.3.3	Сходство белков по структуре . . . . .	16
2.3.4	Сходство лигандов по структуре . . . . .	16
2.3.5	Сходство сайтов связывания . . . . .	17
<b>3</b>	<b>Заключение</b>	<b>19</b>

## Определения, обозначения и сокращения

**Лиганд** – в биологии это химическое соединение, обычно малая молекула, которая образует комплекс с той или иной биомолекулой-мишенью (чаще всего белком) и производит, вследствие такого связывания, те или иные биохимические, физиологические или фармакологические эффекты.

**Мишень** – биомолекула, с которой может связываться молекула-лиганд, производя некоторый биологический эффект в организме.

**FDA** – американское управление по санитарному надзору за качеством пищевых продуктов и медикаментов (англ. Food and Drug Administration) – служба, занимающаяся контролем качества лекарственных препаратов, пищевых продуктов и других продуктов, а также осуществляющая контроль за соблюдением законодательства и стандартов в этой области.

**Метрика подобия или схожести** – способ сопоставить всем парам объектов из входных данных одного типа некоторые безразмерные числа, позволяющие судить о схожести этих объектов.

**а/к** – аминокислотный или аминокислота (в зависимости от контекста).

**ID** – идентификатор.

**БД** – база данных.

**ПМ** – программный модуль.

**GPCR** – рецепторы, сопряженные с G-белком (англ. G-protein-coupled receptors)

**ТМ** – трансмембранный участок

**МГ** – моделирование по гомологии

**МД** – молекулярная динамика

# 1 Введение

Математическое моделирование и вычислительные методы давно стали неотъемлемой частью исследований в биологии и медицине, причем налицо тенденция к росту важности и востребованности таких методов и в фундаментальной науке, и в приложениях. Так, в современной фармакологии высокопроизводительный виртуальный скрининг является важнейшим этапом отсева молекул-кандидатов на статус лекарства перед дорогостоящими клиническими испытаниями. Использование компьютерных расчетов позволяет значительно сократить количество затрачиваемых на поиск лекарства времени и ресурсов.

Одним из методов поиска и отсева кандидатов является сравнение с уже известными лекарствами, имеющими положительные и побочные эффекты. Таким способом можно производить поиск как новых мишеней для уже известных лекарств, так и по известным мишеням находить новые лекарства для них. Кроме поиска лекарства, такие исследования могут помочь в нахождении побочных эффектов и ранее неизвестных эффектов одновременного использования нескольких препаратов. Однако, насколько нам известно, сейчас не существует публично доступного программного модуля, позволяющего искать кандидатов в новые лекарственные средства с использованием широкого спектра различных метрик подобия.

**Целью** данной работы являлось построение такого гибко настраиваемого программного протокола для поиска неспецифичных взаимодействий между лигандами и мишенями.

Для достижения этой цели были предложены следующие этапы:

1. извлечь из доступных баз данных биологических соединений необходимую информацию о подтвержденных FDA лекарствах, соответствующих лигандах и мишенях, и реализовать это в программном коде;
2. создать функции для конвертации данных одной молекулы в требуемые сторонними программами форматы;
3. создать функции для сравнения одного элемента входных данных с

одной подтвержденной FDA записью;

4. реализовать поиск наиболее подходящих в терминах различных метрик подобия лигандов/мишеней/комплексов.

## 1.1 Обзор литературы

### 1.1.1 Полифармакология

При разработке лекарств важно добиться селективности, избавившись от побочных действий. Именно эта парадигма «одно лекарство – одна мишень», так называемая таргетированная терапия, до недавнего времени широко использовалась в фармакологии. С другой стороны, в последнее время стала осознаваться важность полифармакологии, которая означает множественное, но специфичное воздействие лекарства на многие мишени, позволяющее добиться синергетического эффекта и более эффективного лечения комплексных заболеваний, таких как рак[1].

При этом полифармакология может выгодно отличаться от комбинирования нескольких лекарств, так как:

(а) единственная молекула обычно имеет более предсказуемую и безопасную фармакокинетику;

(б) часто действующие на несколько мишеней лекарства имеют большую эффективность на поздних стадиях заболевания;

(в) не нужно учитывать эффекты перекрестного взаимодействия лекарств, которые, являясь негативными, переносятся хуже в случае комбинационной терапии;

(г) при прочих равных меньше вероятность выработки лекарственной устойчивости к одному лекарству, чем к хотя бы одному из набора лекарств [1].

Стоит заметить, что каждый белковый домен в среднем содержит 3-5 связывающих карманов достаточного размера для связывания с типичными малыми лигандами[2]. Таким образом, существует возможность выбрать новый карман, отличный от ранее использовавшихся, для разработки лекарства. К тому же, количество видов связывающих карманов со

статистически значимыми различиями оценивается, как меньшее 400 [2], что позволяет считать полифармакологическую картину взаимодействий лиганд-мишень неизбежной, и потому более перспективной, чем таргетированная.

Новая парадигма подчеркивает важность поиска всевозможных пар взаимодействий лиганд-мишень. Такой анализ может аккумулировать результаты уже известных связей, приводя к построению сложных сетей [1], но важнее уметь предсказывать такие взаимодействия. Так как перебор и оценка силы всех взаимодействий лиганд-мишень *in vivo* и *in vitro* является непрактичной, в этом направлении активно развиваются компьютерные методы [3].

### 1.1.2 Неспецифичные взаимодействия

Неспецифичные взаимодействия – дополнительные взаимодействия выбранного лиганда/мишени с другими, кроме основных, мишенями/лигандами. Одной из основных проблем в поиске таких взаимодействий исходя из структуры является то, что часто эти связи в большой степени определяются подвижными частями рецептора, которые сложно или невозможно исследовать с достаточной атомарной точностью [4].

Принципиально, структурный поиск субъектов неспецифичного взаимодействия может осуществляться по структуре: (а) мишени; (б) лиганда; (в) связывающего сайта [5].

(а) при поиске возможных лигандов по известной структуре мишени, воспроизводится обычный процесс современного дизайна лекарств, а именно высокопроизводительный скрининг по базе возможных лигандов. Таким образом, в сущности, оценивается, насколько сложно найти лиганд для этой мишени. Процесс можно ускорить, используя поиск по так называемым «горячим точкам», то есть набору мест на поверхности мишени, где максимальна энергия связывания с потенциальным лигандом [6], что напоминает концепцию фармакофорного поиска.

(б) поиск по структуре лиганда по сути своей близок к понятию

**repurposing'a**, заключающегося в поиске новых мишеней и применений для лекарств, которые уже выпущены на рынок. Это позволяет сократить расходы на преклиническую стадию и оптимизацию [6; 7].

(в) связывающие сайты могут сравниваться по различным характеристикам, таким как геометрические и физикохимические свойства поверхности мишени, профили взаимодействия или структура остова. Также нахождение связывающих карманов само по себе сложная задача, к которой существует несколько подходов (**добавить, как ее решать**) [8].

Связывающие сайты могут описываться разными способами. Например, как трехмерный граф из вершин-атомов, соединенный ребрами-длинами. Или же как облако точек, то есть чисто геометрически. В этих моделях могут выделяться основанные на фармакофорном принципе черты, которые в дальнейшем позволяют значительно ускорить поиск. Один из наиболее затратных в вычислениях, но и чувствительных методов – построение карт электронной плотности [8].

### 1.1.3 Существующие протоколы

ДОПИСАТЬ ОБЗОР, чего не хватает в существующих, недостатки[9]

## 2 Основное содержание

### 2.1 Общая структура протокола

Опишем общую структуру протокола (см. рис. 1).

Так как для поиска по подобию нужны референсные данные, то необходимо найти их источник. Далее, извлекая сведения об подтвержденных FDA лекарствах, можно, используя эту же и другие базы данных биомолекул, получать необходимую для дальнейшего поиска информацию о свойствах референсных лигандов/мишеней и извлекать структурную информацию о них и их комплексах.

После этого с помощью сторонних программных пакетов и собственного кода достигается цель – нахождение списка лигандов/мишеней/комплексов наиболее близких к входным данным.

### 2.2 Материалы и методы исследования

#### 2.2.1 Биологические базы данных

Описание устройства баз и примеры: Drugbank [10], Uniprot [11], PDB [12], Pubchem [13].

#### 2.2.2 Коэффициенты Жаккара и Танимото

Для удобства работы с мерами сходства как с числами имеет смысл ввести единообразное для самых разных типов данных определение бинарного коэффициента сходства. Коэффициент Жаккара [14] позволяет для любых двух конечных множеств  $A, B$  получить коэффициент подобия  $J(A, B)$  по формуле

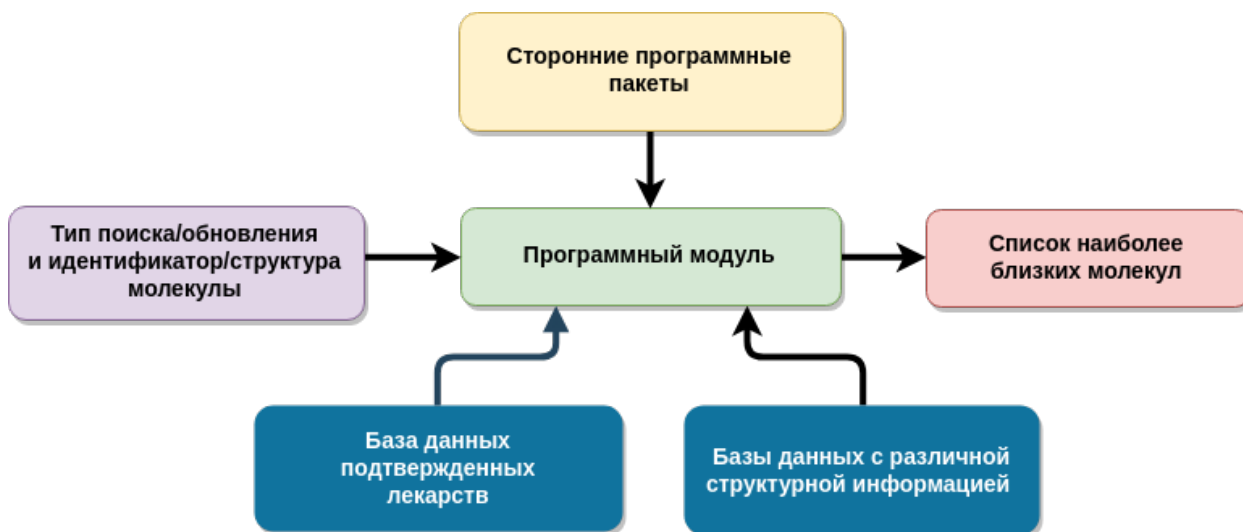
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \in [0, 1]. \quad (1)$$

Видно, что равенство коэффициента нулю означает полное отсутствие сходства между множествами. Чем больше значение  $J(A, B)$ , тем выше степень сходства множеств вплоть до полного совпадения при  $J(A, B) = 1$ .

Частным случаем этого коэффициента является коэффициент Танимото, применяющийся в случае сравнения бинарных множеств. Тогда их



Рис. 1 — Общая схема структуры протокола.



можно охарактеризовать битовыми векторами  $\mathbf{a}, \mathbf{b}$ , а (1) можно переписать проще для прямого вычисления:

$$T(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|^2 + |\mathbf{b}|^2 - \mathbf{a} \cdot \mathbf{b}}. \quad (2)$$

### 2.2.3 Сходство белков по аминокислотной последовательности

Для вычисления сходства белков по их аминокислотным последовательностям был использован модуль Biopython [15]. Он позволяет вычислить подобие и идентичность (**ОПРЕДЕЛЕНИЯ**) двух а/к последовательностей. (**КАК РАБОТАЕТ**) (**ПРИМЕР выравнивания листингом**)

### 2.2.4 Сходство белков по структуре

Для вычисления сходства белков, используя их структуру, использована программа TM-align [16] Сначала с помощью динамического программирования определяется оптимальное наложение двух белков друг на друга, которое, вообще, является NP-трудной задачей [17] и потому требует существенных вычислительных ресурсов. После этого вычисляется TM-score (англ. Template Modeling score)[18]

$$\text{TM-score} = \max \left[ \frac{1}{L_{\text{target}}} \sum_i^{L_{\text{aligned}}} \frac{1}{1 + \left( \frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right], \quad (3)$$

где  $L_{target}$  – количество а/к в целевом белке, сравниваемом с другими;  $L_{aligned}$  – количество а/к в выравненной части двух белков,  $d_i$  – расстояние между  $i$ -й парой остатков в белках;  $d_0$  – нормировочный коэффициент  $d_0(L_{target}) = 1.24\sqrt[3]{L_{target} - 15} - 1.8$ ; сумма производится по парам соответствующих оснований. TM-score нормирован таким образом, что его значение меньше 0,2 говорит об отсутствии корреляций в структурах, а превышение эмпирической границы в 0,5 означает, что укладки белков практически совпадают

Одновременно считается и коэффициент RMSD (англ. Root Mean Square Deviation)

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}, \quad (4)$$

где  $N$  – количество атомов в молекуле;  $\delta_i$  – отклонение  $i$ -го атома от второй структуры.

TM-score считается более подходящим для оценки подобия белков, чем RMSD, так как в отличие от RMSD он не зависит от длин сравниваемых белков (нормируется на полусумму их длин и находится в полуинтервале  $(0, 1]$ ), а также учитывает различные области белков с разными весами, что позволяет получать адекватные результаты для структур с одинаковой глобальной топологией, но небольшими отклонениями по всей длине белка. В случае использования RMSD результат будет подразумевать то, что структуры различны, а TM-score, вероятнее всего, детектирует схожесть

### 2.2.5 Сходство лигандов по структуре

Для нахождения сходства лигандов по их топологической структуре, используются молекулярные отпечатки [19]. Принцип работы заключается в том, что в каждой молекуле находятся и хэшируются заранее заданные топологические пути по связям. (добавить другие типы)Получающиеся битовые строки двух сравниваются, что приводит к коэффициенту Танимото как оценке степени подобия структур.

Мы использовали вычисление таких отпечатков по SMILES и SDF файлу (реализованы соответственно с помощью модулей RDkit [20] и Open Babel [21]).

### 2.2.6 Сходство сайтов связывания

Для вычисления подобия сайтов связывания используется программный пакет IsoMIF [22; 23], считающийся одним из лучших для данного вычисления [8]. (ТОЧНО??? Проверить)

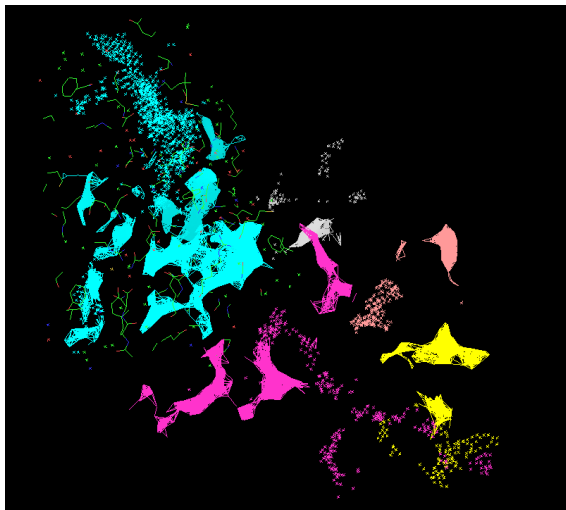
Суть его работы состоит в нахождении полостей в структуре комплекса лиганд-мишень [24; 25] (подробнее как ищутся ).

После этого с помощью встроенной в пакет программы MIF (англ. Molecular Interaction Field) в области сайта связывания строится сетка, в каждом узле которой считаются энергии взаимодействия пробника определенного типа с соседними атомами белка до некоторого радиуса обрезки. Всего используется 6 типов свойств пробника: гидрофобность, ароматичность, способность быть донором/акцептором водородной связи, положительный/отрицательный электрический заряд, энергии спадают одинаково и экспоненциально от расстояния, а значения энергий на 1 Å подбираются эмпирически для всевозможных пар из 6 типов пробника и 13 типов атома белка (см. рис. 2(б)).

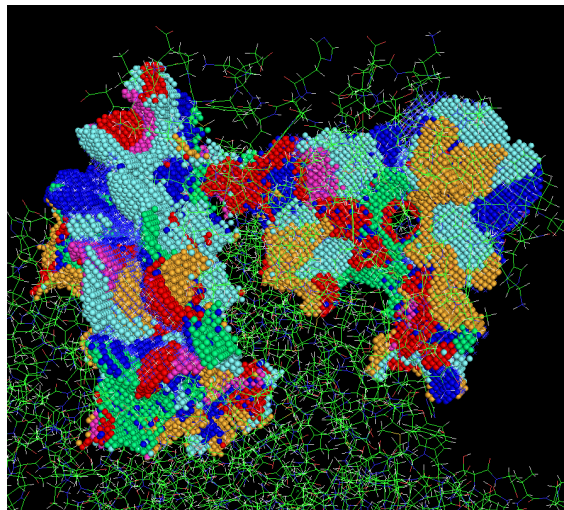
После получения этих сеток для двух сайтов связывания (с 6 значениями энергии в каждом узле) происходит построение графа, у которого вершины обозначают пары узлов сетки, берущихся по одному из сравниваемых структур и имеющих хотя бы один общий из 6 тип энергии больше некоторого порога, то есть является значимым. Ребра в этом графе строятся, если расстояния между соответствующими узлами в двух полостях отличаются менее, чем на 3 Å.

Затем в этом графе производится поиск наибольшей клики (полного подграфа) с помощью алгоритма Брона — Кербоша [26; 27] и с использованием эвристического наблюдения, что наибольшая клика часто находится одной из первых, так что для значительного уменьшения вычислительной

**Рис. 2** — Визуализация работы (а) поиска полостей и (б) вычисления MIF(сделать картинки одного формата )



(а) Пять наибольших найденных с помощью Get Cleft полостей комплекса показаны разными цветами.



(б) Размеченная с помощью MIF полость, разные цвета обозначают тип максимального по энергии взаимодействия в этой области.

нагрузки проводится только 100 поисков по умолчанию.

Результат работы алгоритма – мера подобия полостей, – вычисляется как

$$MSS = \frac{N_c}{N_a + N_b - N_c}, \quad (5)$$

[23]. где  $N_c$  – количество значимых общих типов взаимодействий у всех вершин в клике,  $N_a, N_b$  – количества значимых типов взаимодействий в узлах сеток для первого и второго белка.

TM-align?

## 2.3 Описание программного модуля и результаты

Программный код на языке программирования Python 3.7 [28] и содержащий более 1000 строк доступен в репозитории <https://github.com/antmaxi/E>

Модуль состоит из четырех программ: Search.py, Drugbank.py, IsoMIF.py, Auxiliary.py, различные функции из которых могут вызываться посредством вызова программы Search.py с соответствующими потребностям пользователя или другой программы ключами. **ключи еще не сделаны**

Состав программ:

1. Search.py – все, кроме реализованной с помощью программного пакета IsoMIF, функции обработки данных для получения значений метрик схожести и скрининга по референсным данным.
2. Drugbank.py – функции для извлечения, дополнения и обработки референсных данных из БД Drugbank.
3. IsoMIF.py – функции для проведения поиска схожих комплексов с помощью IsoMIF.
4. Auxiliary.py – вспомогательные функции для работы с файловой системой, для соединения записей об одной молекуле с помощью ID различных БД (Drugbank, PubChem, Uniprot, PDB) и разных характеристик (SMILES, a/a последовательность).

Работа модуля была протестирована на компьютере с операционной системой Ubuntu 16.04 LTS 64-bit, с процессором Intel® Core™ i9-7940X CPU @ 3.10GHz × 28 и видеокартой GeForce GTX 1080 Ti/PCIe/SSE2.

### 2.3.1 Извлечение данных

Было решено производить извлечение референсных данных из базы данных Drugbank, так как она является наиболее полной и хорошо аннотированной БД лекарств сите и содержит не только информацию о лигандах, но и о мишенях вместе со ссылками на другие специфические источники информации.

Хотя в приложении к полной БД Drugbank находятся таблицы с частью нужной для работы протокола информации (ссылки на другие БД, химические идентификаторы), было принято решение извлекать информацию напрямую из полной БД, что позволяет:

- (а) не зависеть от обновления приложений ко всей БД, которые могут запаздывать относительно общего обновления БД;
- (б) делать поиск по нужным записям/свойствам более гибким и простым для будущих модификаций и усовершенствований протокола.

### 2.3.2 Сходство белков по аминокислотной последовательности

В цикле по мишеням, извлеченным из БД Drugbank, производится сравнение а/к последовательности входного белка (можно задать по последовательности или ID Uniprot, тогда а/к последовательность запрашивается у онлайн-сервиса Uniprot) с а/к последовательностями соответствующей мишени. Результатом каждого сравнения является пара чисел (подобие, идентичность), характеризующих степень похожести последовательностей. Затем производится сортировка полученного списка по одной из этих мер подобия, и результат выводится на экран.

Пример использования этой функции в листинге [ref](#). Была взята а/к последовательность GPCR белка родопсина, одного и важнейших для многих областей современной науки. Неудивительно, что наиболее похожей мишенью из БД Drugbank оказался он же сам. Последовательности остальных найденных белков сильно отличаются от данной (падение схожести с 1843 до 318, идентичности с 348 до 156), так что, вероятно, в данном случае сложно говорить о возможности совпадения лигандов родопсина и остальных мишеней из списка.

```
1 In :
2     get_closest_fastas_from_uniprot('P08100', path_to_data_in_fasta, k=0, align_matrix='
   blosum62', sim_or_ident=True)
3 Out :
4
5     position_in_fasta  similarity  identity  sequence  name
6     1392      1392    1843    348    lcl|BSEQ0016346|Rhodopsin
7     1931      1931    318     156    lcl|BSEQ0010278|Cholecystokinin
8     151       151     295     147    lcl|BSEQ0016698|Somatostatin
9     152       152     292     154    lcl|BSEQ0006800|Somatostatin
```

```

10 676      676      270      156      lcl | BSEQ0002303 | Gastrin/cholecystokinin
11 1317     1317     262      139      lcl | BSEQ0010362 | Melatonin
12 711      711      250      147      lcl | BSEQ0001536 | Mu-type
13
14 CPU times: user 10min 41s, sys: 1.26 s, total: 10min 42s
15 Wall time: 10min 43s

```

**Листинг 1** — Сходство мишеней по а/к последовательности с помощью Biopython с входными данными – ID Uniprot белка Родопсина из семейства GPCR.

Сравнение одной пары занимает от долей секунды до около 10 секунд, вся база данных порядка 10-30 минут.

### 2.3.3 Сходство белков по структуре

Список ID PDB, в которые входит белок, берется с помощью онлайн-сервиса БД Uniprot. Далее необходимо из этих структур выбрать те, в которых белок находится сам по себе, без других аминокислотных цепей (так как TM-align считает схожесть для всех атомов в аминокислотах, даже если белков или их частей в файле несколько)

Для такой фильтрации списка используется ранее упоминавшееся вычисление схожести а/к последовательностей. .... **В ПРОЦЕССЕ, КОД ПИШУ**

### 2.3.4 Сходство лигандов по структуре

**ОТНОСИТЕЛЬНО ПОДРОБНО РАСПИСАНО, ТАК ВСЕ БУДУТ ТИПЫ ПОИСКОВ (+КАРТИНКИ НАДО)**

Поиск по ID SMILES:

Из БД Drugbank извлекаются соответствующие ID SMILES подтвержденных FDA лигандов. Информация дополняется извлеченными из БД PubChem (<https://pubchem.ncbi.nlm.nih.gov>) идентификаторами тех лигандов, у которых нет ID SMILES в БД Drugbank, но имеющих запись в БД PubChem.

Далее, с помощью ПМ RDkit строятся и сравниваются молекулярные отпечатки, вычисляется коэффициент Танимото между ними, приво-

для отсортированного по степени подобия списку подтвержденных FDA лигандов.

Поиск по структуре в формате SDF:

С сайта БД Drugbank скачивается файл со структурами всех лигандов в формате SDF, из него по ID в БД Drugbank извлекаются структуры подтвержденных FDA лигандов. Затем по этим структурам получаются молекулярные отпечатки, вычисляется коэффициент Танимото между ними, результируя в искомом списке.

Пример работы:

```
1 In:
2 get_closest_ligands_from_3d_structure(path_to_structure, path_to_sdf_approved, root,
3 fptype='maccs', number_to_print=5)
4 Out:
5 Name          Tanimoto coeff Drugbank ID Fingerprint_type
6 0 Dichlorobenzyl alcohol      0.343284 DB13269          fp2
7 1          Tiludronic acid    0.317647 DB01133          fp2
8 2          Chloroxylenol      0.308824 DB11121          fp2
9 3          Sulconazole        0.290598 DB06820          fp2
10 4          Guanabenz          0.268293 DB00629          fp2
11 CPU times: user 1.87 s, sys: 7.68 ms, total: 1.88 s
12 Wall time: 1.88 s
```

**Листинг 2** — Сходство лигандов по молекулярным отпечаткам с помощью Open Babel для входных данных – SDF структуры молекулы ...

НУЖНО ЛИ ПОДРОБНО ОБЪЯСНЯТЬ, ЧТО ВВОДИТСЯ, ЧТО ВЫВОДИТСЯ? КАРТИНКА СО СРАВНЕНИЕМ ИЗОБРАЖЕНИЙ МОЛЕКУЛ

Время поиска по фармакофорам 1-3 сек.

### 2.3.5 Сходство сайтов связывания

Для скрининга комплекса из входных данных по референсным комплексам необходимо каким-то образом получить их структуры. Для этого для лиганда и мишени строятся два списка ID PDB, включающих в себя их.

Для мишени список составляется просто по ID Uniprot с помощью запросов к онлайн-сервису БД Uniprot, позволяющего находить соответствующие данному белку ID PDB (при этом может случиться, что в этих



структурах находится белок как сам по себе, так и в комплексе с некоторым лигандом).

Для лиганда по его SMILES ищутся ID PDB, в которых есть похожая на него структура. Для этого нужно ввести требуемый уровень коэффициента Танимото или же указать шаг, с которым этот коэффициент будет уменьшаться с 1 до тех пор, пока не будет найдена хотя бы одна структура в БД PDB.

Затем находятся общие элементы списков ID PDB для мишени и лиганда. Из них можно взять тот, с которым подобие лиганда максимально для ускоренного поиска, или же использовать все.

В конце концов, с помощью ПМ IsoMIF находятся коэффициент подобия карманов связывания для всех найденных референсных структур комплексов и входного комплекса, и выводятся наиболее похожие.

Для одной пары полостей время построения MIF обычно около 1-3 мин, время вычисления IsoMIF около 3-5 мин, остальные части этого поиска длятся пренебрежимо мало.

Возможно, в будущем для ускорения работы стоит сделать параллельную версию этой части протокола.

### 3 Заключение

Построен протокол поиска подобных и соответствующих подтвержденным FDA лигандов/мишеней/комплексов по нескольким типам входных данных с использованием различных способов задания вычисления подобия: по 1D-, 2D-, 3D-структурам. Работа протокола проверена и задокументирована, программный код и примеры опубликованы в открытом доступе. (В ПРОЦЕССЕ )

Из-за своей модульной структуры в будущем протокол может дополняться новыми вариантами нахождения меры схожести. Также на его основе с применением машинного обучения могут быть построены более сложные методы поиска, учитывающие несколько различных метрик подобия.

## Список литературы

1. *Anighoro A., Bajorath J., Rastelli G.* Polypharmacology: Challenges and opportunities in drug discovery // *Journal of Medicinal Chemistry*. — 2014. — Т. 57, № 19. — С. 7874–7887. — ISSN 15204804. — DOI: 10.1021/jm5006463.
2. Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function / J. Skolnick, M. Gao [и др.] // *Bioorganic & Medicinal Chemistry Letters*. — 2015. — Март. — Т. 25, № 6. — С. 1163–1170. — DOI: 10.1016/j.bmcl.2015.01.059. — URL: <https://doi.org/10.1016/j.bmcl.2015.01.059>.
3. Computational polypharmacology: a new paradigm for drug discovery / R. Chaudhari [и др.] // *Expert Opinion on Drug Discovery*. — 2017. — Т. 12, № 3. — С. 279–291. — DOI: 10.1080/17460441.2017.1280024. — eprint: <https://doi.org/10.1080/17460441.2017.1280024>. — URL: <https://doi.org/10.1080/17460441.2017.1280024> ; PMID: 28067061.
4. *Loving K. A., Lin A., Cheng A. C.* Structure-Based Druggability Assessment of the Mammalian Structural Proteome with Inclusion of Light Protein Flexibility // *PLOS Computational Biology*. — 2014. — Июль. — Т. 10, № 7. — С. 1–13. — DOI: 10.1371/journal.pcbi.1003741. — URL: <https://doi.org/10.1371/journal.pcbi.1003741>.
5. *Rognan D.* Structure-Based Approaches to Target Fishing and Ligand Profiling // *Molecular Informatics*. — 2010. — Март. — Т. 29, № 3. — С. 176–187. — DOI: 10.1002/minf.200900081. — URL: <https://doi.org/10.1002/minf.200900081>.
6. Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery / D. R. Hall, D. Kozakov, A. Whitty, S. Vajda // *Trends in Pharmacological Sciences*. — 2015. — Ноябрь. — Т. 36, № 11. — С. 724–736. — ISSN 0165-6147. — DOI: 10.1016/j.tips.2015.08.003. — URL: <https://doi.org/10.1016/j.tips.2015.08.003>.
7. On the integration of in silico drug design methods for drug repurposing / E. March-Vila, L. Pinzi [и др.] // *Frontiers in Pharmacology*. — 2017. — Т. 8, МАЙ. — С. 1–7. — ISSN 16639812. — DOI: 10.3389/fphar.2017.00298. — arXiv: arXiv:1011.1669v3.
8. *Ehrt C., Brinkjost T., Koch O.* Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design // *Journal of Medicinal Chemistry*. — 2016. — Май. — Т. 59, № 9. — С. 4121–4151. — ISSN 0022-2623. — DOI: 10.1021/acs.jmedchem.6b00078. — URL: <https://doi.org/10.1021/acs.jmedchem.6b00078>.
9. Large-scale detection of drug off-targets: Hypotheses for drug repurposing and understanding side-effects / M. Chartier, L. P. Morency, M. I. Zylber, R. J. Najmanovich // *BMC Pharmacology and Toxicology*. — 2017. — Т. 18, № 1. — С. 1–16. — ISSN 20506511. — DOI: 10.1186/s40360-017-0128-7.
10. URL: <https://www.drugbank.ca>.
11. URL: <https://www.uniprot.org>.
12. URL: <https://www.rcsb.org>.
13. URL: <https://pubchem.ncbi.nlm.nih.gov>.
14. *Jaccard P.* Comparative de la distribution florale dans une portion des Alpes et des Jura // *Bulletin de la Société Vaudoise des Sciences Naturelles*. — 1901. — № 7. — С. 547–579.
15. URL: <https://biopython.org>.
16. URL: <https://zhanglab.ccmb.med.umich.edu/TM-align>.

17. *Lathrop R. H.* The protein threading problem with sequence amino acid interaction preferences is NP-complete // "Protein Engineering, Design and Selection". — 1994. — Т. 7, № 9. — С. 1059–1068. — DOI: 10.1093/protein/7.9.1059. — URL: <https://doi.org/10.1093/protein/7.9.1059>.
18. *Levitt M., Gerstein M.* A unified statistical framework for sequence comparison and structure comparison // Proceedings of the National Academy of Sciences. — 1998. — Май. — Т. 95, № 11. — С. 5913–5920. — DOI: 10.1073/pnas.95.11.5913. — URL: <https://doi.org/10.1073/pnas.95.11.5913>.
19. Molecular fingerprint similarity search in virtual screening / A. Cereto-Massagué, M. J. Ojeda [и др.] // Methods. — 2015. — Т. 71, № C. — С. 58–63. — ISSN 10959130. — DOI: 10.1016/j.ymeth.2014.08.005.
20. URL: <http://www.rdkit.org/docs/index.html>.
21. URL: [http://openbabel.org/wiki/Main\\_Page](http://openbabel.org/wiki/Main_Page).
22. URL: <https://github.com/mtthchrtr/IsoMif>.
23. *Chartier M., Najmanovich R.* Detection of Binding Site Molecular Interaction Field Similarities // Journal of Chemical Information and Modeling. — 2015. — Т. 55, № 8. — С. 1600–1615. — ISSN 15205142. — DOI: 10.1021/acs.jcim.5b00333.
24. *Gaudreault F., Morency L.-P., Najmanovich R. J.* NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID // Bioinformatics. — 2015. — Абр. — btv458. — DOI: 10.1093/bioinformatics/btv458. — URL: <https://doi.org/10.1093/bioinformatics/btv458>.
25. *Laskowski R. A.* SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions // Journal of Molecular Graphics. — 1995. — Окт. — Т. 13, № 5. — С. 323–330. — DOI: 10.1016/0263-7855(95)00073-9. — URL: [https://doi.org/10.1016/0263-7855\(95\)00073-9](https://doi.org/10.1016/0263-7855(95)00073-9).
26. *Bron C., Kerbosch J.* Algorithm 457: finding all cliques of an undirected graph // Communications of the ACM. — 1973. — Т. 16, № 9. — С. 575–577. — ISSN 00010782. — DOI: 10.1145/362342.362367.
27. *Tomita E., Tanaka A., Takahashi H.* The worst-case time complexity for generating all maximal cliques and computational experiments // Theoretical Computer Science. — 2006. — Т. 363, № 1. — С. 28–42. — ISSN 03043975. — DOI: 10.1016/j.tcs.2006.06.015.
28. URL: <https://www.python.org/downloads/release/python-370>.