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 референсных лигандов 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

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 категорий товаров, 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 — аминокислота.

АКП — аминокислотная последовательность.

ID — идентификатор.

БД — база данных.

ПМ — программный модуль.

ПП — программный пакет.

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 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 лиганды для них. Кроме поиска лекарства, такие исследования могут помочь в нахождении побочных эффектов и ранее неизвестных последствий одновременного использования нескольких препаратов. Однако, насколько нам известно, сейчас не существует публично доступного программного модуля или пакета, 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 по референсным данным наиболее подходящих в терминах различных метрик подобия для 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 б**о**льшую 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 вероятность привыкания к единственному лекарству меньше, 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, по оценке, не превышает 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 и их поиск

Неспецифичные взаимодействия — дополнительные взаимодействия выбранного лиганда/мишени с другими, кроме основных, мишенями/лигандами. Одной из главных 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, обычно воспроизводится процесс современного дизайна лекарств, а именно высокопроизводительный скрининг по базе возможных лигандов. Таким образом, в сущности, оценивается, насколько сложно найти лиганд для этой мишени.

Схожесть мишеней можно оценивать по их АКП (1D – мера) или по структурным особенностям (3D – меры). Вычисление 3D – мер часто бывает ресурсоемким, но этот процесс можно ускорить, используя поиск по так называемым «горячим точкам», то есть набору мест на поверхности мише-

ни, где максимальна энергия связывания с потенциальным лигандом [6]. Это напоминает концепцию фармакофорного поиска, то есть нахождения определенных пространственных и электронных структур, особенно энергетически выгодных для связывания лиганда.

(б) Поиск по структуре лиганда близок по своей сути к понятию лекарственной репозиции, заключающемуся в поиске новых мишеней и применений для лекарств, которые уже выпущены на рынок. Это позволяет сократить расходы на преклиническую стадию и оптимизацию [6; 7].

Для нахождения сходства лигандов по их топологической структуре могут использоваться молекулярные отпечатки нескольких типов. Существуют разнообразные принципы построения отпечатков: могут хэшироваться различные топологические пути, могут искаться фармакофоры или определенные подструктуры, существуют и способы с использованием только текста SMILES [8]. В получающихся для двух лигандов битовых строках вычисляется относительное количество общих битов – свойств, что и приводит к оценке степени подобия структур (см. коэффициент Танимото в разделе 2.2.2).

Преимуществом молекулярных отпечатков является относительная быстрота их построения и возможность предварительного их вычисления для всех элементов референсных данных. После этого для сравнения с новым элементом достаточно лишь раз посчитать отпечаток для него, а также пройти по базе данных (БД) для вычисления коэффициентов Танимото, что принципиально является просто вычислением бинарных «И» и «ИЛИ» для битовых строк. Однако, для эффективной работы принципа молекулярных отпечатков нужно среди всего их множества выбирать наиболее подходящие к конкретному классу сравниваемых соединений [8].

(в) Связывающие карманы могут сравниваться по различным характеристикам, таким как геометрические и физико-химические свойства поверхности мишени, профили взаимодействия или структура остова [9]. Соответственно, точность и ресурсоемкость различных методов варьируются в широких диапазонах.

Полости могут описываться разными способами. Например, как трехмерный граф из вершин-атомов, соединенных 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 [9].

Кроме того, нахождение полостей само по себе сложная задача, к которой существует несколько подходов. Подавляющее большинство способов нахождения полостей подразумевает модель твердых шаров для структуры молекул, что сводит задачу преимущественно в область обработки геометрических данных [10], хотя часто используются и вычисления пробных потенциалов с построением соответствующих карт и дальнейшим выделением областей методами вычислительной математики (например, методом Нелдера — Мида).

Существуют методы, основанные на построении сетки, которые благодаря совершенствованию вычислительных возможностей в последнее время стали вновь широко использоваться. Их преимущество в практически неограниченном росте точности при уменьшении размера сетки. Недостаток – в «прямоугольности» и дискретности сетки, которые, впрочем, с помощью различных ухищрений могут сглаживаться.

Другой класс методов основан на диаграммах Вороного и триангуляциях Делоне [11], которые позволяют автоматически избавиться от дискретности и сильной зависимости точности построения от ресурсоемкости, а также более применимы для анализа структур в динамике. Недостаток этого класса состоит в неуниверсальности замены атомов твердыми шарами, а также сложности обработки построенных диаграмм.

Еще один класс методов использует, по сути, множественное «встраивание» пробника в структуру. В простейшем случае пробником является сфера, ее радиус может уменьшаться до достижения нужной точности. Преимущество этого класса в возможности автоматически учитывать раз-

мер лиганда, задавая размер пробника порядка размера лиганда. Проблема такого рода методов в сильной зависимости от размера молекул и недостаточной точности [10].

Комбинация различных методов из вышеперечисленных позволяют использовать их преимущества при уменьшении эффекта недостатков.

1.1.3 Актуальность

Большинство существующих работ по поиску неспецифичных взаимодействий для построения соответствующих предсказаний рассматривают преимущественно один метод вычисления подобия. [12; 13] Это мотивирует нас на создание протокола, способного использовать несколько различных типов вычисления подобия молекул/комплексов и гибкого к добавлению новых способов.

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. Таким источником могут являться биологические базы данных, содержащие необходимые для оценок сходства входных данных с референсными. Извлекая сведения об подтвержденных FDA лекарствах, можно, используя эту же и другие базы данных биомолекул, получать и сохранять в структурированном виде необходимые для дальнейшего поиска сведения о свойствах референсных лигандов/мишеней и извлекать структурную информацию о них и их комплексах.

После этого с помощью сторонних программных модулей и пакетов для оценки сходства молекул/комплексов и обработки данных, а также собственного модуля достигается цель — нахождение списка лигандов/мишеней/комплексов наиболее близких к входным данным.

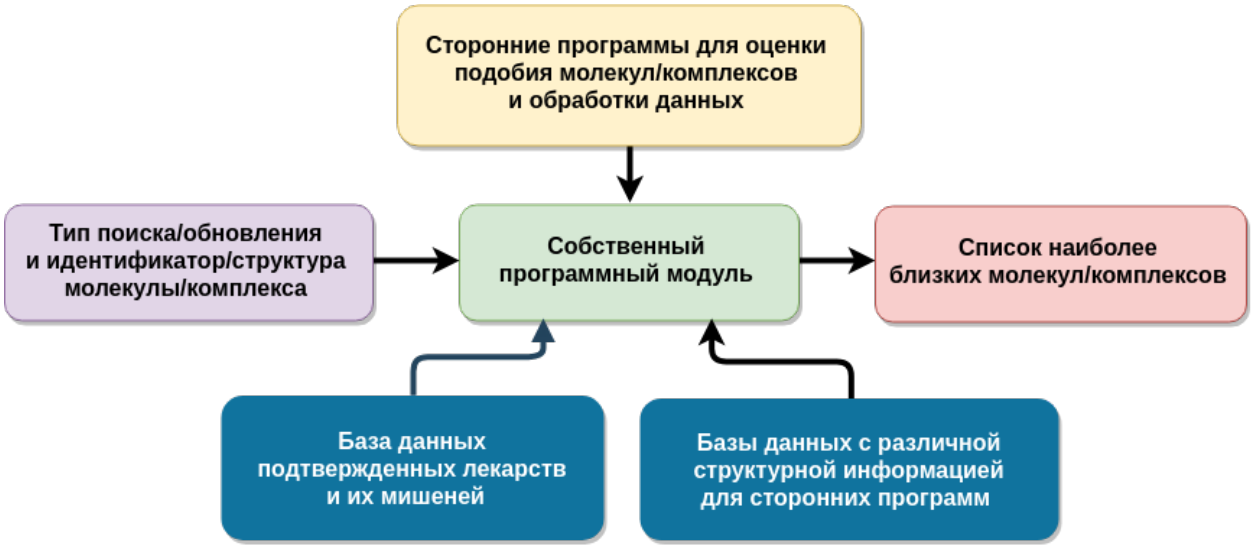

Рисунок 1 — Общая схема структуры протокола.

2.2 Материалы и методы исследования

2.2.1 Биологические и химические базы данных

Объем и качество биологических данных растут с каждым годом, что требует эффективных способов их хранения и оперирования ими. С этой целью были созданы, создаются и обновляются биологические базы данных, содержащие самую разнообразную информацию о биологических и биоактивных молекулах, тканях, видах, лекарствах и так далее. Обычно биологические БД имеют свой сайт и реализованный пользовательский интерфейс/API для извлечения информации об отдельных записях и поиску по набору свойств. Записи могут аннотироваться и проверяться автоматически и вручную, от чего часто зависит качество представляемых в биологических БД сведений.

Примерами таких БД являются:

- Drugbank [14] — специализируется на подтвержденных FDA и экспериментальных лекарствах;
- Uniprot [15] — содержит информацию о белках и протеомах;
- PDB [16] — состоит из записей о структурной информации биомолекул и комплексов;
- PubChem [17] — является более общей химической БД, однако содержит химическую информацию и о биомолекулах, биоактивных веществах. При поиске абсолютно новых лекарств они часто еще не включены в класс биоактивных веществ, так что общие химические БД также используются в биологических исследованиях и приложениях.

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 коэффициент сходства. Коэффициент Жаккара [18] 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 A, B. Тогда их можно охарактеризовать битовыми векторами $\mathbf{a}, \mathbf{b}$, и в этом случае $T(\mathbf{a},\mathbf{b}) \equiv J(A,B)$ из (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 двух АКП обычно используется алгоритм BLAST [19] или его вариации. Принцип работы алгоритма заключается в поиске выравниваний двух АКП, возможно с пробелами, таким образом, чтобы суммарное сходство S, определяемое как сумма весов, взятых из предопределенной матрицы замен (см. рис. 2) для всех выравненных пар а/к и пробелов из двух последовательностей:

$$S = \sum_{i=1}^{L_{align}} M\dot{S}ub,$$

где сумма ведется по всем номерам аминокислот/пробелов в выравнивании, M - матрица замен с учетом замен пробелов/на пробелы, Sub - матрица количеств замен каждого типа в данном выравнивании. кривовато Вначале строятся частичные выравнивания с помощью поиска соответствий в 3 а/к, затем пробуется продлить эти соответствия сначала без пробелов, потом с ними. В итоге, вычисляется коэффициент суммарного сходства и идентичность, определяемая как количество пар одинаковых а/а в выравнивании, а лучшее выравнивание выводится на экран.

Для вычисления сходства белков описанным выше образом по их аминокислотным последовательностям был использован модуль Biopython [20].

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Рисунок 2 — Наиболее часто используемая в выравниваниях АКП матрица замен BLOSUM62.

2.2.4 Сходство белков по структуре

Чтобы вычислить сходство белков, используя их структуру, использована программа TM-align [21; 22]. Сначала определяется оптимальное наложение двух белков друг на друга, нахождение которого, вообще, является NP-трудной задачей без точного решения [23], и поэтому обычно требует существенных вычислительных ресурсов. Однако, применяемый в TM-align алгоритм позволяет уменьшить время работы на порядок без потери точности, что и послужило причиной его выбора в качестве основы для этой части протокола.

Для выравнивания структур сначала с помощью динамического программирования определяется выравнивание вторичных структур и строится соответствующая бинарная оценочная матрица. Потом меньший белок «протягивается» по большему для беззазорного выравнивания и вычисляется соответствующая оценочная матрица. Матрица для итогового выравнивания берется как средняя первых двух, а выравнивание строится посредством алгоритма динамического программирования. В итоге вычис-

ляется TM-score (англ. Template Modeling score)[24; 25].

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

где «max» обозначает наилучшее значение после пространственного наложения, 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; 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 не зависит от длин сравниваемых белков. Также из статистических исследований следует, что значение TM-score меньше 0,2 говорит об отсутствии корреляций в структурах, а превышение эмпирической границы в 0,5 означает, что укладки белков практически совпадают [22; 26].

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

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i}, \tag{4}$$

где N — количество атомов в первой структуре; δ_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 [22].

2.2.5 Сходство лигандов по структуре

Мы использовали вычисление молекулярных отпечатков по SMILES (текстовый отпечаток [27]) и SDF файлу (топологические пути и поиск подструктур [8]). Нахождение этих отпечатков реализовано с помощью модулей RDkit [28] и Open Babel [29] соответственно. Не представляет сложности при необходимости добавить другие типы и реализации вычисления молекулярных отпечатков, например фармакофорные отпечатки из тех же модулей.

2.2.6 Сходство сайтов связывания

Для вычисления подобия сайтов связывания используется программный пакет (ПП) IsoMIF [30; 31], считающийся одним из лучших для данного вычисления, хотя и вычислительно затратным] [9].

Суть его работы состоит в нахождении и обработке полостей структуры комплекса лиганд-мишень[32] Для нахождения пустот используется реализация алгоритма SURFNET [33] (см. рис 3). Этот алгоритм не позволяет выделить именно полости между несколькими молекулами, потому что также помечает внутренние пустоты и углубления в молекулах как полости. При этом поиск производится начиная с наибольших полостей к меньшим, а в большинстве случаев именно наибольшая полость в структуре комплекса и является сайтом связывания с лигандом [33].

Принцип нахождения полостей базируется на постепенном вписывании сфер наибольшего радиуса посередине между всеми парами атомов, где каждый из пары берется из отдельной молекулы. Атомы также считаются шарами, с которыми новые сферы не должны пересекаться. Сферы вписываются в структуру до тех пор, пока нельзя будет вписать еще одну минимального заранее заданного радиуса. После этого в центре каждой сферы строится гауссиан, вносящий вклад в общую карту интенсивности, а по суммарной карте интенсивности, из которой по заданному порогу легко выделить поверхность и полости [33].

После этого с помощью встроенной в пакет программы MIF (англ.

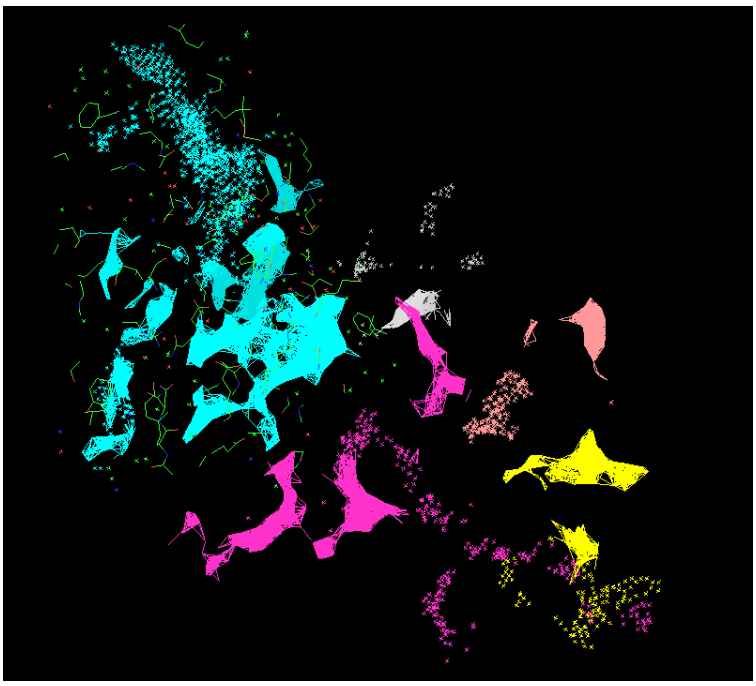

Рисунок 3 — Пять наибольших полостей в структуре 1E8X из БД PDB, найденных с помощью Get Cleft, показаны разными цветами.

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 видов атома белка (см. рис. 4).

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 типов 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

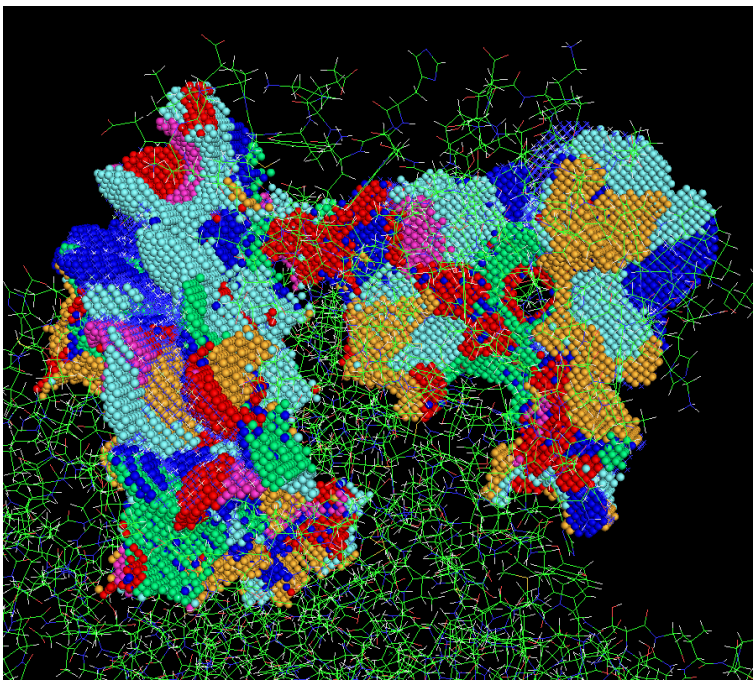

Рисунок 4 — Размеченная посредством MIF наибольшая полость в структуре 1E8X из БД PDB, разные цвета обозначают тип максимального по энергии взаимодействия в этой области.

подграфа) с помощью алгоритма Брона — Кербоша [34; 35] и с использованием эвристического наблюдения, что наибольшая клика часто находится одной из первых, так что для значительного уменьшения вычислительной нагрузки проводится только 100 поисков по умолчанию.

Результат работы алгоритма — мера подобия полостей, — вычисляется как

$$\mathrm{MSS} = \frac{N_c}{N_a + N_b - N_c}, \tag{5}$$

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 [31].

Также в протокол включена возможность использовать TM-score для сравнения структур комплексов. Такая функция в программе TM-align появилась недавно и описание ее работы еще не опубликовано, но, вероятно, она работает по тому же принципу, что и соответствующее сравнение белков (см. раздел 2.2.4).

2.3 Описание программного модуля и результаты

Программный код на языке программирования Python 3.7 [36] и содержащий около 1600 строк обновлять доступен в открытом репозитории по адресу https://github.com/antmaxi/BSc_thesis.

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 ПП 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 ПП 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, а/а 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 БД Drugbank, так как она является одной из наиболее полных и хорошо аннотированных БД лекарств [37] и содержит информацию не только о лигандах, но и об их 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.

При извлечении лигандов с информацией о них считалось, что лиганд подтвержден FDA, если среди указанных в БД лекарственных продуктов с его участием есть хотя бы один подтвержденный FDA. Это значит, что эта молекула в некоторых условиях безвредна для человека, так что можно включить ее в предварительный поиск лекарств.

Поиск производился обработкой скачанной полной БД Drugbank в формате XML с помощью последовательного выбора потомков по нужным именам полей, начиная с корней всей БД. Использование сторонних модулей обработки без прямой итерации по элементам невозможно из-за большого (1,3 Гб) размера БД. Из-за этого поиск производится сравнительно дольше (3–4 минуты).

После нахождения нужной информации она сохраняется на диск компьютера, чтобы в дальнейшем можно было не просматривать всю БД при каждом запуске поиска по коэффициент подобия. Всего найдено 3220 различных лигандов, 2513 их разных мишеней и 9048 связей лиганд-мишень.

В будущем поиск такого рода можно усовершенствовать путем приоритетного рассмотрения тех пар лиганд-мишень, которые входят в подтвержденные FDA лекарства. Так же можно будет учитывать известные побочные эффекты референсных лекарств при построении прогноза перспективности входных данных для дальнейшего изучения.

2.3.2 Сходство белков по аминокислотной последовательности

В цикле по мишеням, извлеченным из БД Drugbank, производится сравнение АКП входного белка (можно задать по последовательности или ID Uniprot, тогда АКП запрашивается у онлайн – сервиса Uniprot) с АКП соответствующей мишени. Результатом каждого сравнения является пара чисел (подобие, идентичность), характеризующих степень похожести АКП. Затем производится сортировка полученного списка по одной из этих мер подобия, и результат выводится на экран. Причем для наиболее сходных молекул строится выравнивание для указания основных областей схожести/различия белков.

Пример использования этой функции приведен в листинге 1 . Была взята АКП GPCR человеческого белка родопсина, одного из важнейших белков у человека и ключевого инструмента оптогенетики. Неудивительно, что наиболее похожей мишенью из БД Drugbank оказался он же сам. Наилучшие найденные белки, как и следовало ожидать, также принадлежат семейству GPCR, но их а/к последовательности достаточно сильно отличаются от данной (падение схожести с 1843 до 318, идентичности с 348 до 156).

```
In:
  get_closest_fastas_from_uniprot('P08100', path_to_data_in_fasta, k=0, align_matrix='
    blosum62', sim_or_ident=True)
Out:

  similarity    identity          name
  1392    1843  348   lcl|BSEQ0016346|Rhodopsin
  1931    318   156   lcl|BSEQ0010278|Cholecystokinin_receptor_type_A
  151     295   147   lcl|BSEQ0016698|Somatostatin_receptor_type_5
  152     292   154   lcl|BSEQ0006800|Somatostatin_receptor_type_2
  676     270   156   lcl|BSEQ0002303|Gastrin/cholecystokinin_type_B_receptor
  1317    262   139   lcl|BSEQ0010362|Melatonin_receptor_type_1A

  ...

  Name = lcl|BSEQ0010278|Cholecystokinin_receptor_type_A
  Similarity=318.5, identity=156
  Matrix blosum62, number of alignments = 1
  MNGTEGPNFYVPFSNATGVVRSPFEYPQYY-LAEP-----WQFS---

  MLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLGGEIAL
  WSLVVLAIERYVVVCKPM-SNFFGENHAIMGVAFTWVMALACAAP-PLAGWSRYIP--
    EGLQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIIIFFCYGQLVFTV----------KEAAAQQQESATT------------QK
    ---------------------------------AEKEVTRMVIIMVIAFLICWVPYASV---
  -AFYIFTHQGSNFG-PIFMTIPAFFAKSAAIYNPVIYIMMNKQFRNCMLTTICCGKNP--------LGDDEASATVSKT--------
```

```
    ETSQVAPA
  |....  ...|..||.|.......|..... |..|     ||..   .|....|||.|||
    |.|...|....|..||..|..||.|||.||...|......|...|...|.||...|.....|..........||....|||
  ...|||. |.......||....|.||........| |.  .|...|
    ............|...|...|........|.||.|.....||......          |..|.....|.|.            ||
                                  |.|.|.||.|..|..|..||.|..|.    |..........|
    ||.........|..  ||.||..|||.||.....|..|..||        .|..|...|....        ....|.|.
  MDVVD--SLLVNGSNITPPCELGLENETLFCLDQPRPSKEWQPAVQILLYSLIFLLSVLG---
    NTLVITVLIRNKRMRTVTNIFLLSLAVSDLMLCLFCMPFNLIPNLLKDFIFGSAVCKTTTYFMGTSVSVSTFNLVAISLERYGAICKP
  LQSRVWQTKSHALKVIAATWCLSFTIMTPYPI--
    YSNLVPFTKNNNQTANMCRFLLPNDVMQQSWHTFLLLILFLIPGIVMMVAYGLISLELYQGIKFEASQKKSAKERKPSTTSSGKY
  EDSDGCYLQKTRPPRKLELRQLSTGSSSRANRIRSNSSAANLMAKKRVIRMLIVIVVLFFLCWMPIFSANAWRAYDTASAERRLSGT
  PISFILLLSYTSSCV--NPIIYCFMNKRFRLGFMATFPCCPNPGPPGARGEVGEEEEGGTTGASLSRFSYSHMSASVPPQ
  Score=318.5

  ...

CPU times: user 10min 41s, sys: 1.26 s, total: 10min 42s
Wall time: 10min 43s
```

Листинг 1 — Определение сходства мишеней по а/к последовательности посредством Biopython, входные данные — ID Uniptor человеческого родопсина из семейства GPCR.

Делать прямые выводы о наличии/отсутствии общих лигандов только из выравнивания сложно, это требует отдельного поиска и анализа карманов связывания (в случае GPCR чаще всего он известен), однако наличие достаточной схожести дает основания исследовать этот вопрос для этой пары мишеней глубже. Также то, что найденные белки принадлежат тому же семейству GPCR, говорит о том, что работа функции подтверждает корректность работы функции.

Сравнение одной пары занимает от долей секунды до около 10 секунд, сравнение со всей БД порядка 10-30 минут.

2.3.3 Сходство белков по структуре

Список ID PDB, в которые входит белок, автоматически берется с помощью онлайн – сервиса БД Unirprot. Далее необходимо из этих структур выбрать или выделить те, в которых белок находится сам по себе, без других полипептидных цепей (так как TM-align считает схожесть для всех атомов в аминокислотах, и если белков или их частей в файле несколько, то будет обрабатывать только первую встреченную цепь; так что этот алгоритм применим только для определенного класса структур)

Для такой фильтрации списка используется ранее упоминавшееся вычисление схожести АКП с помощью ПМ Biopython. Среди встречающихся в файле цепей ищется и выделяется посредством ПМ Biopython та, которая имеет наибольшую степень 1D –подобия с последовательностью входного белка.

После предварительной обработки всех извлеченных из БД Drugbank белков, а также протеомов человека, крысы и мыши, искомый белок с помощью TM-align сравнивается с выделенными структурами для одного из упомянутых наборов белков. Они сортируются по вычисленным TM-score и выводятся на экран. В ПРОЦЕССЕ, КОД ПОЧТИ ГОТОВ

2.3.4 Сходство лигандов по структуре

КАРТИНКИ

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, приводя к отсортированному по степени подобия списку подтвержденных FDA лигандов. Пример приведен в листинге 2.

```
In:
  get_closest_smiles_names('ClCCNC(=O)N(CCCl)N=O', root, 5)
Out:
                query                      smiles        similarity       name
  124   O=NN(CCCl)C(=O)NCCCl      ClCCNC(=O)N(CCCl)N=O    1.000000    Carmustine
  919   O=NN(CCCl)C(=O)NCCCl  ClCCN(N=O)C(=O)NC1CCCCC1    0.644351     Lomustine
  465   O=NN(CCCl)C(=O)NCCCl            NCC=C.ClCC1CO1    0.550000     Sevelamer
  1110  O=NN(CCCl)C(=O)NCCCl              NC(CO)(CO)CO    0.540984  Tromethamine
  1066  O=NN(CCCl)C(=O)NCCCl                 CCCCCON=O    0.524590  Amyl Nitrite
CPU times: user 2.66 s, sys: 7.99 ms, total: 2.67 s
```

```
Wall time: 2.66 s
```

Листинг 2 — Сходство лигандов по текстовым молекулярным отпечаткам с помощью ПМ RDkit для входных данных — SMILES структуры молекулы.

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 приведен в листинге 3.

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

Листинг 3 — Сходство лигандов по топологическим молекулярным отпечаткам с помощью ПМ Open Babel для входных данных — SDF структуры молекулы.

НУЖНО ЛИ ПОДРОБНО ОБЪЯСНЯТЬ, ЧТО ВВОДИТСЯ, ЧТО ВЫВОДИТСЯ?

КАРТИНКА СО СРАВНЕНИЕМ ИЗОБРАЖЕНИЙ МОЛЕКУЛ

Время поиска по молекулярным отпечатком для всей БД около 1-3 сек.

2.3.5 Сходство сайтов связывания

Для сравнения комплекса из входных данных с референсными комплексами необходимо каким-то образом получить их структуры. С этой целью 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 (у 1939 из 3220 извлеченных лигандов SMILES имеется) ищутся и сохраняются для будущих запусков его ID PDB, в которых есть похожая на него структура. Для этого нужно ввести достаточный уровень коэффициент Танимото или же указать шаг, с которым этот коэффициент будет уменьшаться начиная с 1 до тех пор, пока не будет найдена хотя бы одна структура в БД PDB или не будет достигнуть задаваемый максимальный уровень похожести. При фиксированном шаге в 0.05 время поиска всех PDB (определяется главным образом временем получения ответа с сайта БД Uniprot) занимает для различных максимальных уровней: 1.0 (то есть поиск точных совпадений) – около 30 минут, 0.9 – около 40 минут. С фиксированным уровнем подобия результаты: 0.9 – около 35 минут, 0.8 – около 45 минут. Стоит заметить, что также существуют таблицы, в которых эти списки уже составлены, но для некоторых заданных уровней подобия. Нас же подход позволяет гибко настраивать параметры поиска в зависимости от требуемой точности и располагаемых ресурсов.

Затем находятся и также общие элементы списков ID PDB для мишени и лиганда.

Корректность работы этой части была проверена взятием случайного лиганда (ацетазоламида) и одной и его мишеней (карбоангидразы 4). Три наиболее близкие лиганду общие структуры в PDB и их уровни подобия лигандов совпадают с данными в приложенной к БД Drugbank таблице. Однако, в этой таблице указаны только 3 лучшие структуры по сходству к лиганду, а наш метод позволяет найти нужное нам количество с необходимой точностью, так что здесь, как и с первичным поиском подтвержденных лигандов, стратегия самостоятельного поиска вместо простой обработки приложений к БД дает свои плоды.

Далее, простым циклом по референсным парам лиганд-мишень можно получить для каждого комплекса коэффициент его сходства с исходным комплексом и затем отсортировать результат по убыванию подобия.

В качестве сторонних программ, вычисляющих коэффициенты схожести, используются ПП IsoMIF и TM-align. Находятся коэффициенты подобия карманов связывания для всех найденных референсных структур комплексов и входного комплекса, и выводятся наиболее похожие.

ПРИМЕР

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. По всей БД поиск произведен не был из-за слишком большого времени работы. Предлагается в дальнейшем усовершенствовать протокол добавлением параллельных версий наиболее ресурсоемких и времязатратных частей.

Для одной пары полостей время работы TM-align составляет несколько секунд, для всей БД Сколько.

3 Заключение

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

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. Усовершенствование протокола может быть дополнительно осуществлено учетом данных о побочных эффектах известных лекарств, которые могут влиять на предсказания эффектов найденных кандидатов.

Список использованных источников

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. — Нояб. — 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. — М. — Т. 8. — С. 1—7. — ISSN 16639812. — DOI: 10.3389/fphar.2017.00298. — arXiv: arXiv:1011.1669v3.
8. Molecular fingerprint similarity search in virtual screening / A. Cereto-Massagué, M. J. Ojeda [и др.] // Methods. — 2015. — Т. 71, № С. — С. 58—63. — ISSN 10959130. — DOI: 10.1016/j.ymeth.2014.08.005.
9. *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.
10. Visual Analysis of Biomolecular Cavities: State of the Art / M. Krone, B. Kozlikova [и др.] // Computer Graphics Forum. — 2016. — Июнь. — Т. 35, № 3. — С. 527—551. — DOI: 10.1111/cgf.12928. — URL: https://doi.org/10.1111/cgf.12928.
11. *Aurenhammer F.* Voronoi diagrams—a survey of a fundamental geometric data structure // ACM Computing Surveys. — 1991. — Сент. — Т. 23, № 3. — С. 345—405. — DOI: 10.1145/116873.116880. — URL: https://doi.org/10.1145/116873.116880.
12. A comprehensive map of molecular drug targets / R. Santos, O. Ursu [и др.] // Nature Reviews Drug Discovery. — 2016. — Дек. — Т. 16, № 1. — С. 19—34. — DOI: 10.1038/nrd.2016.230. — URL: https://doi.org/10.1038/nrd.2016.230.

13. 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.

14. URL: https://www.drugbank.ca.

15. URL: https://www.uniprot.org.

16. URL: https://www.rcsb.org.

17. URL: https://pubchem.ncbi.nlm.nih.gov.

18. *Jaccard P.* Comparative de la distribuition florale dans une portion des Alpes et des Jura // Bulletin de la Société Vaudoise des Sciences Naturelles. — 1901. — № 7. — С. 547—579.

19. Basic local alignment search tool / S. F. Altschul, W. Gish [и др.] // Journal of Molecular Biology. — 1990. — Т. 215, № 3. — С. 403—410. — ISSN 0022-2836. — DOI: https://doi.org/10.1016/S0022-2836(05)80360-2. — URL: http://www.sciencedirect.com/science/article/pii/S0022283605803602.

20. URL: https://biopython.org.

21. URL: https://zhanglab.ccmb.med.umich.edu/TM-align.

22. *Zhang Y.*, *Skolnick J.* TM-align: a protein structure alignment algorithm based on the TM-score // Nucleic Acids Research. — 2005. — Янв. — Т. 33, № 7. — С. 2302—2309. — ISSN 0305-1048. — DOI: 10.1093/nar/gki524. — eprint: http://oup.prod.sis.lan/nar/article-pdf/33/7/2302/7127128/gki524.pdf. — URL: https://doi.org/10.1093/nar/gki524.

23. *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.

24. *Zhang Y.*, *Skolnick J.* Scoring function for automated assessment of protein structure template quality // Proteins: Structure, Function, and Bioinformatics. — 2004. — Т. 57, № 4. — С. 702—710. — DOI: 10.1002/prot.20264. — URL: https://doi.org/10.1002/prot.20264.

25. *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.

26. *Xu J.*, *Zhang Y.* How significant is a protein structure similarity with TM-score = 0.5? // Bioinformatics. — 2010. — Февр. — Т. 26, № 7. — С. 889—895. — DOI: 10.1093/bioinformatics/btq066. — URL: https://doi.org/10.1093/bioinformatics/btq066.

27. *Weininger D.*, *Weininger A.*, *Weininger J. L.* SMILES. 2. Algorithm for generation of unique SMILES notation // Journal of Chemical Information and Modeling. — 1989. — Май. — Т. 29, № 2. — С. 97—101. — DOI: 10.1021/ci00062a008. — URL: https://doi.org/10.1021/ci00062a008.

28. URL: http://www.rdkit.org/docs/index.html.

29. URL: http://openbabel.org/wiki/Main_Page.

30. URL: https://github.com/mtthchrtr/IsoMif.

31. *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.

32. *Gaudreault F.*, *Morency L.-P.*, *Najmanovich R. J.* NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID // Bioinformatics. — 2015. — Авг. — btv458. — DOI: 10.1093/bioinformatics/btv458. — URL: https://doi.org/10.1093/bioinformatics/btv458.

33. *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.

34. *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.

35. *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.

36. URL: https://www.python.org/downloads/release/python-370.

37. DrugBank 5.0: a major update to the DrugBank database for 2018 / D. S. Wishart, Y. D. Feunang [и др.] // Nucleic Acids Research. — 2017. — Нояб. — Т. 46, № D1. — С. D1074—D1082. — DOI: 10.1093/nar/gkx1037. — URL: https://doi.org/10.1093/nar/gkx1037.