Министерство образования и науки РФ
Московский Физико-Технический Институт
(национальный исследовательский университет)
Факультет общей и прикладной физики
Кафедра вычислительной физики конденсированного состояния и живых систем

Протокол для поиска возможных неспецифичных взаимодействий лигандов

Диплом на соискание степени бакалавра

Выполнил:
студент группы 527и
Максимов Антон Сергеевич

Научный руководитель:
(к. ф.-м. н.)
Попов Пётр Анатольевич

Долгопрудный 2019

Аннотация

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 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 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 скрининг является важнейшим этапом отсева молекул-кандидатов на статус лекарства перед дорогостоящими клиническими испытаниями, позволяющим значительно сократить количество затрачиваемых на поиск лекарства времени и сил.

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 важным мишеням находить новые лекарства. Однако, насколько нам известно, на сегодняшний день не существует публично доступного программного модуля, позволяющего искать кандидатов в новые лекарственные средства с помощью различных способов вычисления 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 предсказумую 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 меньше вероятность вырабатывания 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 [**?**]. 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 [**?**], 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 направлениии активно развиваются компьютерные методы [2].

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 [3].

Принципиально, структурный поиск субъектов неспецифичного взаимодействия может осуществляться по структуре: (а) мишени; (б) лиганда; (в) связывающего сайта [4].

(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 [5], что напоминает концепцию фармакофорного поиска.

(б) поиск по структуре лиганда по сути своей близок к понятию **repurposing'a**, которое заключается в поиске новых мишеней и применений для лекарств, которые уже выпущены на рынок. Это позволяет сократить расходы на преклиническую стадию и оптимизацию [5, 6].

(в) связывающие сайты могут сравниваться по различным характе-

ристикам, 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) [7].

Связывающие сайты могут описываться разными способами. Например, как трехмерный граф из вершин-атомов, содиненный 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 [7]. [8] [9] [10]

1.1.3 Существующие протоколы

ДОПИСАТЬ ОБЗОР, чего не хватает в существующих, недостатки[11]

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, Uniprot, PDB, Pubchem.

2.2.2 Коэффициент Танимото

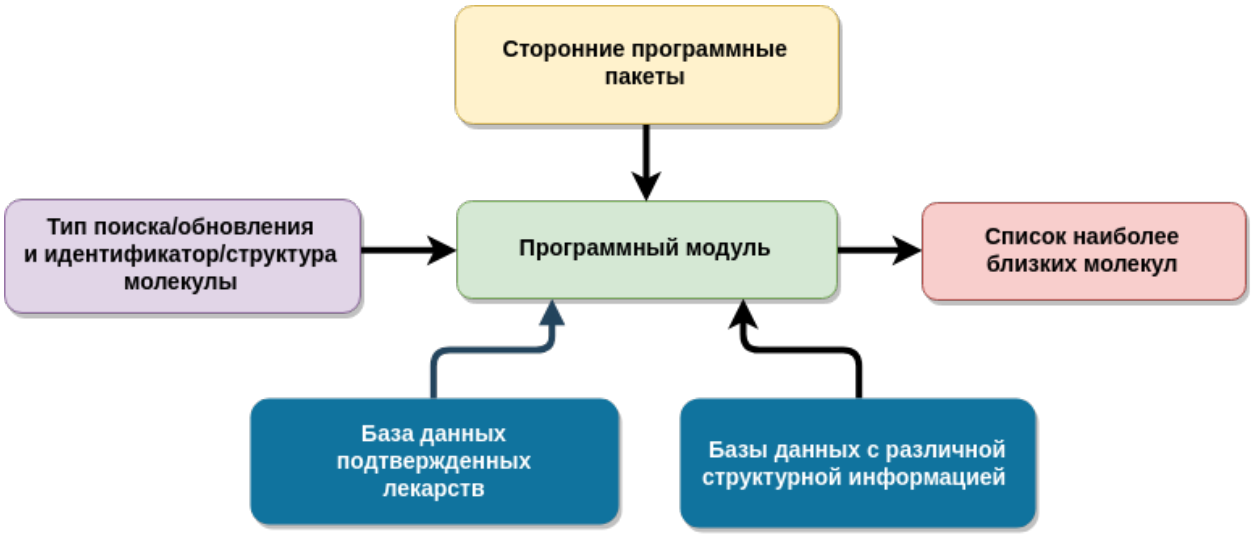
2.2.3 Сходство белков по аминокислотной последовательности

Чтобы вычислить сходство белков по их аминокислотным последовательностям был использован модуль Biopython(https://biopython.org). Он позволяет вычислить подобие и идентичность (ОПРЕДЕЛЕНИЯ) двух а/к последовательностей.(КАК РАБОТАЕТ)

2.2.4 Сходство белков по структуре

Для вычисления сходства белков, используя их структуру, использована программа TM-align (https://zhanglab.ccmb.med.umich.edu/TM-align) [12]. Сначала с помощью динамического программирования определяется

Рис. 1 — Общая схема структуры протокола.



оптимальное наложение двух белков друг на друга, которое, вообще, является NP-трудной задачей [] и потому требует существенных вычислительных ресурсов. После этого вычисляется TM-score (англ. Template Modeling score)[13]

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

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 соответствущих 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 [12].

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

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i}, \tag{2}$$

где N – количество атомов в молекуле; 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 [12].

2.2.5 Сходство лигандов по структуре

Чтобы найти сходство лигандов по их топологической структуре, ичспользуются молекулярные отпечатки [15]. 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 модуля RDkit (http://www.rdkit.org/docs/index.html) и Open Babel (http://openbabel.org/wiki/Main_Page))

2.2.6 Сходство сайтов связывания

Для вычисления подобия сайтов связывания используется программный пакет IsoMIF (https://github.com/mtthchrtr/IsoMif)[16], считающийся одним из лучших для данного вычисления [7]. (ТОЧНО??? Проверить)

Суть его работы состоит в нахождении полостей в структуре компекса лиганд-мишень(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: гидробофность, 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 [17, 18] 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.

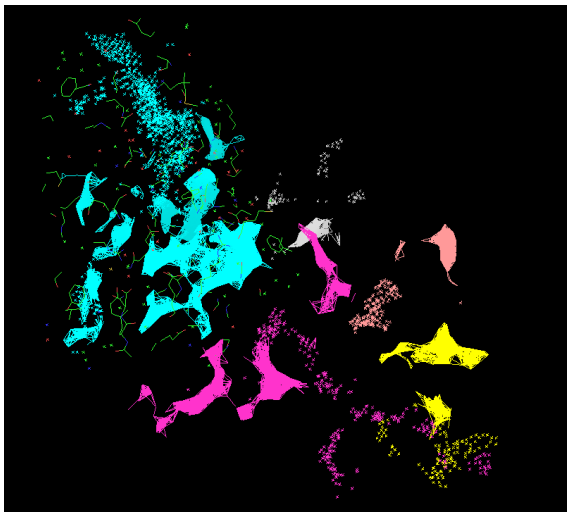
Результат работы алгоритма – мера подобия полостей, – вычисляется как

$$\mathrm{MSS} = \frac{N_c}{N_a + N_b - N_c}, \tag{3}$$

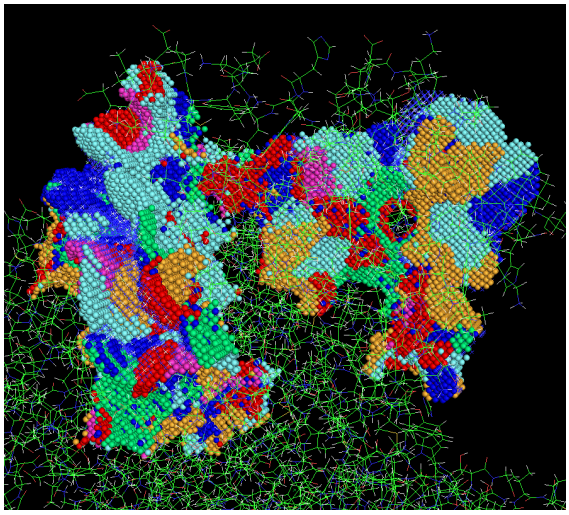
[16]. где N_c – количество значимых общих типов взаимодействий у всех вершин в клике, N_a, N_b – количества значимых типов взаимодействий в узлах сеток для первого и второго белка.

TM-align?

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



(a) Найденные полости комплекса показаны разными цветами.



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

2.3 Описание программного модуля и результаты

Программный код на языке программирования Python 3.7 (https://www.python.org/downloads/release/python-370) и содержащий более ?1000 строк доступен в репозитории 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.

В Search.py находится большее количество функций обработки данных для получения значений метрик схожести. В Работа модуля была протестирована но компьютере с операционной системой Ubuntu 16.04 LTS 64-bit, с процессором Intel® Core^{TM} 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 (https://www.drugbank.ca/), так как она является наиболее полной и хорошо аннотированной БД лекарств **??** и содержит не только информацию о лигандах, но и о мишенях вместе с 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 зависить от обновления приложений ко всей БД. которые могут запаздывать относительно общего обновления БД;

(б) делать поиск по нужным записям/свойства более гибким и простым для будущих модификаций и усовершенствований протокола.

2.3.2 Сходство белков по аминокислотной последовательности

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

2.3.3 Сходство белков по структуре

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 их нет в БД Drugbank, но которые имеют 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.

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:

```
In:
  get_closest_ligands_from_3d_structure(path_to_structure, path_to_sdf_approved, root,
                                        fptype='maccs', number_to_print=5)
  (ЗАМЕНЮ, КОГДА СДЕЛАЮ ЧЕРЕЗ argparse)
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 1 — Сходство лигандов по молекулярным отпечаткам с помощью Open Babel

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. поля каждой пары лиганд-мишень, извлеченных из БД Drugbank,

Время построения MIF 1-3 мин, вычисления IsoMIF 3-5 мин, остальные части длятся пренебрежимо мало.

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.

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. (В ПРОЦЕССЕ)

Наверное, про GPCR не стоит говорить отдельно, они же нигде мной не выделяются (то, что писал осенью, не нужно)

3.1 Рецепторы, сопряженные с G-белком

Семейство GPCR состоит из около 800 многофункциональных белков-рецепторов [19], регулирующих разнообразные внутриклеточные сигнальные каскады в ответ на гормоны, нейротрансмиттеры, ионы, фотоны, одоранты и другие стимулы[20]. Поэтому они играют важнейшую роль в физиологии и разработке лекарств, представляя собой привлекательную мишень для лекарственных средств. Около трети всех лекарств, одобренных Управлением по санитарному надзору за качеством пищевых продуктов и медикаментов США, действуют именно на мишени из этого класса, хотя это всего чуть более ста различных рецепторов, то есть порядка десятой части всего семейства[21].

Белки, принадлежащие семейству GPCR, состоят из семи трансмембранных спиралей, связанных тремя внутриклеточными и внеклеточными петлями. Внеклеточная часть, с которой связывается лиганд, также включает в себя N-конец. Внутриклеточная часть содержит кроме петель также восьмую спираль и C-конец и взаимодействует с G белками, аррестинами

и другими **downstream effectors** [22].

Трансмембранная часть является наиболее консервативной в структуре белков семейства, что не мешает рецепторам из различных подсемейств обеспечивать крайнее разнообразие в форме, размере и электростатических свойствах связывающихся с ними лигандов за счет вариаций в структуре связывающих карманов. Их гидрофобность и закрытость с внеклеточной стороны связаны с функциями рецептора [22].

Несмотря на колоссальный прогресс в кристаллографии GPCR, пространственная структура известна только у небольшой доли рецепторов. Важная роль компьютерного моделирования состоит в раскрытии структур остальных рецепторов и их комплексов [23] и последующем рациональном создании лекарств.

3.2 Сайты связывания и специфичность

Хотя природные лиганды внутри семейства GPCR очень разнообразны, белки одного подтипа имеют практически одинаковые конформации активных сайтов, что позволяет моделировать их компьютерными методами с высокой точностью [24]. Некоторые подсемейства рецепторов взаимодействуют с одним и тем же **endogenous** лигандом, и в этом случае отсутствие больших различий в ортостерических связывающих карманах, где и происходит взаимодействие **endogenous** лиганда и рецептора, представляет вызов для поиска селективных лигандов и является одной из главных проблем в разработке безопасных и эффективных лекарств, действующих на GPCR [25].

Исторически дизайн лекарств был направлен на создание лигандов, подобных **endogenous**, которые искусственно активировали сигнальные пути. Другой классический подход состоит в создании антагонистов – веществ, способных конкурировать с природным лигандом, не активируя при этом работу рецептора.

В последнее время произошел взрывной рост количества новых методов использования GPCR в качестве мишени. Например, многие недавно

разработанные лиганды действуют в активном сайте, топологически отделенном от ортостерического. Такие сайты и лиганды называются «аллостерическими», причем лиганды могут как усиливать работу рецептора, так и ослаблять ее [26]. Возможна даже комбинация фармакофоров ортостерических и аллостерических лигандов (**bitopic ligands**) в одном лиганде, который будут теоретически иметь лучшую аффинность и селективность за счет большего количества связей с рецептором[27].

Простейший механизм активации GPCR включает в себя два состояния, между которыми балансирует рецептор: активное, в котором происходит передача сигнала внутрь клетки, и неактивное. Агонисты смещают это равновесие в сторону активной конформации, обратные агонисты - в сторону неактивного. Эффективность агониста определяется преобладанием *аффинности к активному состоянию над неактивным(звучит коряво)*. К настоящему моменту стало очевидно, что различные лиганды, воздействуя на один и тот же рецептор, могут стабилизировать его в различных конформациях, так что активными остается только часть всех возможных сигнальных путей, в который вовлечен этот рецептор. Такой процесс называется **biased** агонизм [28].

Дальше всё по моделированию по гомологии (с осени), что, наверное, другая тема и не будет использоваться здесь (тем более, что MODELLER не использую, часть с мутациями, кажется, выходит за пределы)

3.3 Построение модели белка

Итогом развития геномного секвенирования в последнее время стал резкий рост количества известных белковых последовательностей, в то время, как только около одной сотой доли последовательностей охарактеризована с атомистической точностью и с использованием экспериментальных методов определения структуры [29].

В таких условиях полученные компьютерными методами модели структур белков часто являются ценными для выдвижения проверяемых гипотез. Такие модели, в целом, создаются с использованием методов сравни-

тельного моделирования или свободного моделирования (**free modelling**), также называемых «ab initio» или «de novo»[29].

Сравнительное моделирование, или моделирование на основе гомологии (МГ), базируется на построении модели по известным структурам близких (**related** белков, как по шаблонам. Принцип **Free modelling** подхода в использовании не структуры близких белков, а в применении разнообразных методов, комбинирующих физику и известные особенности структур белков, например, сопоставление большого количества небольших фрагментов, выделенных из известных структур белков. Конструирование белков этим методом обычно чрезвычайно затратно в плане вычислений[29]. Современные пакеты моделирования зачастую комбинируют эти два подхода, используя, если доступны шаблоны, МГ для построения основы-скелета белка, а затем уточняя положения петель, боковых цепей и частей без шаблона.

Моделирование требует наличия схемы сэмплирования конформаций с целью получения набора альтернативных структур. Также необходима оценивающая функция для ранжирования этих конформаций по качеству. Для этих целей было предложено большое количество физически обоснованных (**physics-based**) функций энергии и статистических потенциалов, полученных из анализа известных структур [30].

Так, например, MODELLER сначала накладывает входную последовательность на шаблонный остов, а затем, внося случайные смещения атомов, ищет локальный минимум оценивающей функции, повторяя эту процедуру несколько раз [29].

Значительное увеличение количества решенных структур GPCR позволяет строить с атомистической точностью пространственные структуры большого количества рецепторов при отсутствии экспериментально полученной структуры целевого белка [31].

Построение модели состоит из нескольких этапов, результатом которых является физически и биологически адекватная модель белка. Ими являются: (1) определение сходства набора последотельностей с целевой

и выделение шаблонов; (2) выравнивание целевой последовательности и шаблона(ов); (3) построение модели, основанной на выравнивании с выбранными шаблонами; (4) предсказание точности модели [29].

3.3.1 Выделение шаблонов и множественное выравнивание последовательностей

Сначала, беря аминокислотную последовательность заданного белка с неизвестной пространственной структурой, необходимо получить набор белков с известной 3D-структурой для последующего «сшивания» соответствующих участков.

Для этого используются различные подходы к поиску шаблонов и множественному выравниванию соответствующих им последовательностей:

1) PSI-BLAST[32, 33]: с помощью матрицы вероятностей замен аминокислот BLOSUM62 и штрафа за пропуски производится поиск наиболее близких последовательностей в базе до некоторого порога близости. Далее, с использованием программы Clustal [34] эти последовательности выравниваются с исходной в совокупности. А именно, ... **алгоритм**

2) HHBlits [35] – быстрее и чувствительнее BLAST, но ... (**недостатки**). Использует предварительно кластеризованные с помощью kClustal [36] базы Uniprot и nr.

Общее описание алгоритма: (a) приведение запроса из одной последовательности (выравнивание) или нескольких (множественное выравнивание) к скрытой марковской модели (CMM, сопоставление каждой позиции в последовательности вектора вероятностей размерности 1x20 обнаружить ту или иную аминокислоту на этом месте) производится добавлением к исходной последовательностей, отличающихся от нее заменами аминокислот на похожие по физико-химическим свойствам. При этом учитывается локальный контекст в 13 оснований вокруг замены, а суммарная энергия этих замен должна быть меньше некоторого задаваемого порога, что важно для скорости и чувствительности алгоритма. Далее, каждая строка в исходной CMM причисляется к одному из 219 типичных профилей-кластеров

последовательностей, создавая строку-профиль этой СММ. После этого сначала проводится выравнивание этого профиля СММ с предварительно созданными профилями из базы данных, потом – обычное выравнивание уже входной последовательности внутри соответствующего профилю СММ кластера базы данных. (**наверное, так много тут не надо, просто разбирался в алгоритме**)

3) GPCRdb.org – при исследовании GPCR проще всего использовать готовые выравнивания с сайта базы данных GPCRdb (*зачем тогда вообще другие нам?*).

3.3.2 Моделирование белка и оценка качества модели

Моделирование из выровненного набора шаблонов может производиться при помощи различных программных пакетов для предсказания четвертичной структуры белков. Среди них наиболее известным является MODELLER, но и, например, Rosetta [37] [38], Itasser[39], RaptorX[40] могут быть использованы. [41]

При использовании пакета MODELLER в случае идентичности последовательностей более 30%, в среднем более $\sim$ 60 % скелетных атомов моделируются корректно со средним квадратическим отклонением позиций $C\alpha$ атомов менее $3,5$ Å. При меньшей идентичности, как правило, результат хуже [29].

какие у разных программ особенности и различия?

Оценка качества модели [42],

3.3.3 Уточнение модели

Хотя современные программы МГ производят проверки физической и химической адекватности полученных моделей, за счет неточностей в выбранном силовом поле и просто недостаточности исходного покрытия последовательностями шаблонами, итоговые модели могут иметь значительные отличия от реальной структуры.

Улучшение точности модели может производиться средствами моле-

кулярной динамики (МД) [43]. Производя множественные траектории МД с полученной МГ моделью, помещенной в нативную среду, могут быть получены несколько устойчивых подсостояний белка с локально минимальными свободными энергиями, которые и будут считаться наиболее вероятными конформациями белка в мембране, что важно для GPCR.

3.4 Поиск неспецифичных взаимодействий

Определение, где находится активный центр и взаимодействует ли лиганд с рецептором, может быть проведено различными методами. Возможно использование докинга [44] и основанного на машинном обучении поиска....(*что еще?*)

Заметки:
слишком много слов «взаимодействие», надо придумать синонимов

Список литературы

[1] Andrew Anighoro, J??rgen Bajorath, and Giulio Rastelli. Polypharmacology: Challenges and opportunities in drug discovery. *Journal of Medicinal Chemistry*, 57(19):7874–7887, 2014.

[2] Rajan Chaudhari et al. Computational polypharmacology: a new paradigm for drug discovery. *Expert Opinion on Drug Discovery*, 12(3):279–291, 2017. PMID: 28067061.

[3] Kathryn A. Loving, Andy Lin, and Alan C. Cheng. Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility. *PLOS Computational Biology*, 10(7):1–13, 07 2014.

[4] Didier Rognan. Structure-based approaches to target fishing and ligand profiling. *Molecular Informatics*, 29(3):176–187, mar 2010.

[5] David R. Hall, Dima Kozakov, Adrian Whitty, and Sandor Vajda. Lessons from hot spot analysis for fragment-based drug discovery. *Trends in Pharmacological Sciences*, 36(11):724–736, Nov 2015.

[6] Eric March-Vila, Luca Pinzi, Noé Sturm, Annachiara Tinivella, Ola Engkvist, Hongming Chen, and Giulio Rastelli. On the integration of in silico drug design methods for drug repurposing. *Frontiers in Pharmacology*, 8(MAY):1–7, 2017.

[7] Christiane Ehrt, Tobias Brinkjost, and Oliver Koch. Impact of binding site comparisons on medicinal chemistry and rational molecular design. *Journal of Medicinal Chemistry*, 59(9):4121–4151, May 2016.

[8] Alvaro CortΓ©s-Cabrera, Garrett M Morris, Paul W Finn, Antonio Morreale, and Federico Gago. Comparison of ultra-fast 2d and 3d ligand and target descriptors for side effect prediction and network analysis in polypharmacology. *British Journal of Pharmacology*, 170(3):557–567, 2013.

[9] Michal Brylinski. Local alignment of ligand binding sites in proteins for polypharmacology and drug repositioning. In *Protein function prediction*,

volume 1611 of *Methods in Molecular Biology*, pages 109 – 122. Humana Press, New York, NY, 2017.

[10] Rajiv Gandhi Govindaraj and Michal Brylinski. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinformatics*, 19(1):91, Mar 2018.

[11] Matthieu Chartier, Louis Philippe Morency, María Inés Zylber, and Rafael J. Najmanovich. Large-scale detection of drug off-targets: Hypotheses for drug repurposing and understanding side-effects. *BMC Pharmacology and Toxicology*, 18(1):1–16, 2017.

[12] Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 01 2005.

[13] M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences*, 95(11):5913–5920, May 1998.

[14] M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences*, 95(11):5913–5920, 2002.

[15] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C):58–63, 2015.

[16] Matthieu Chartier and Rafael Najmanovich. Detection of Binding Site Molecular Interaction Field Similarities. *Journal of Chemical Information and Modeling*, 55(8):1600–1615, 2015.

[17] Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[18] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42, 2006.

[19] Robert Fredriksson, Malin C. Lagerström, Lars-Gustav Lundin, and Helgi B. Schiöth. The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, 63(6):1256–1272, 2003.

[20] Daniel Hilger, Matthieu Masureel, and Brian K. Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural and Molecular Biology*, 25(1):4–12, 2018.

[21] Alexander S. Hauser, Misty M. Attwood, Mathias Rask-Andersen, Helgi B. Schiöth, and David E. Gloriam. Trends in gpcr drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery*, 16:829 EP –, Oct 2017.

[22] Vsevolod Katritch, Vadim Cherezov, and Raymond C. Stevens. Diversity and modularity of G protein-coupled receptor structures. *Trends in Pharmacological Sciences*, 33(1):17–27, 2012.

[23] Irina Kufareva, Vsevolod Katritch, RaymondB C. Stevens, and Ruben Abagyan. Advances in gpcr modeling evaluated by the gpcr dock 2013 assessment: Meeting new challenges. *Structure*, 22(8):1120 – 1139, 2014.

[24] Vsevolod Katritch, Vadim Cherezov, and Raymond C. Stevens. Structure-function of the g proteinвЂ“coupled receptor superfamily. *Annual Review of Pharmacology and Toxicology*, 53(1):531–556, 2013. PMID: 23140243.

[25] Vsevolod Katritch, Irina Kufareva, and Ruben Abagyan. Structure based prediction of subtype-selectivity for adenosine receptor antagonists. *Neuropharmacology*, 60(1):108 – 115, 2011. High Resolution.

[26] Jeremy Shonberg, Ralf C. Kling, Peter Gmeiner, and Stefan LΓ¶ber. Gpcr crystal structures: Medicinal chemistry in the pocket. *Bioorganic & Medicinal Chemistry*, 23(14):3880 – 3906, 2015. Selective GCPR Ligands.

[27] Denise Wootten, Arthur Christopoulos, and Patrick M. Sexton. Emerging paradigms in gpcr allostery: implications for drug discovery. *Nature Reviews Drug Discovery*, 12:630 EP –, Aug 2013. Review Article.

[28] J. Robert Lane, Lauren T. May, Robert G. Parton, Patrick M. Sexton, and Arthur Christopoulos. A kinetic view of gpcr allostery and biased agonism. *Nature Chemical Biology*, 13:929 EP –, Aug 2017. Perspective.

[29] Benjamin Webb and Andrej Sali. Protein structure modeling with MODELLER. In *Methods in Molecular Biology*, pages 39–54. Springer New York, 2017.

[30] Guang Qiang Dong, Hao Fan, Dina Schneidman-Duhovny, Ben Webb, and Andrej Sali. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics*, 29(24):3158–3166, 2013.

[31] Christofer S. Tautermann. Gpcr homology model generation for lead optimization. In *Methods in Molecular Biology*, pages 115–131. Springer New York, nov 2017.

[32] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[33] Stephen F. Altschul, Thomas L. Madden, Alejandro A. SchΓ¤ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[34] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[35] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature Methods*, 9:173 EP –, Dec 2011.

[36] Maria Hauser, Christian E Mayer, and Johannes Söding. kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics*, 14(1):248, 2013.

[37] Carol A. Rohl, Charlie E.M. Strauss, Kira M.S. Misura, and David Baker. Protein structure prediction using rosetta. In *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 66 – 93. Academic Press, 2004.

[38] Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. High-resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742, oct 2013.

[39] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The i-TASSER suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, jan 2015.

[40] Sheng Wang, Wei Li, Shiwang Liu, and Jinbo Xu. Raptorx-property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44(W1):W430–W435, 2016.

[41] Haiyou Deng, Ya Jia, and Yang Zhang. Protein structure prediction. *International Journal of Modern Physics B*, 32(18):1840009, 2018.

[42] Min-Yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–2524, Nov 2006. 17075131[pmid].

[43] Amin Nowroozi and Mohsen Shahlaei. A coupling of homology modeling with multiple molecular dynamics simulation for identifying representative conformation of gpcr structures: a case study on human bombesin receptor subtype-3. *Journal of Biomolecular Structure and Dynamics*, 35(2):250–272, 2017. PMID: 26922838.

[44] H. Pradeep and G. K. Rajanikant. A rational approach to selective pharmacophore designing: an innovative strategy for specific recognition of gsk3β. *Molecular Diversity*, 16(3):553–562, Aug 2012.