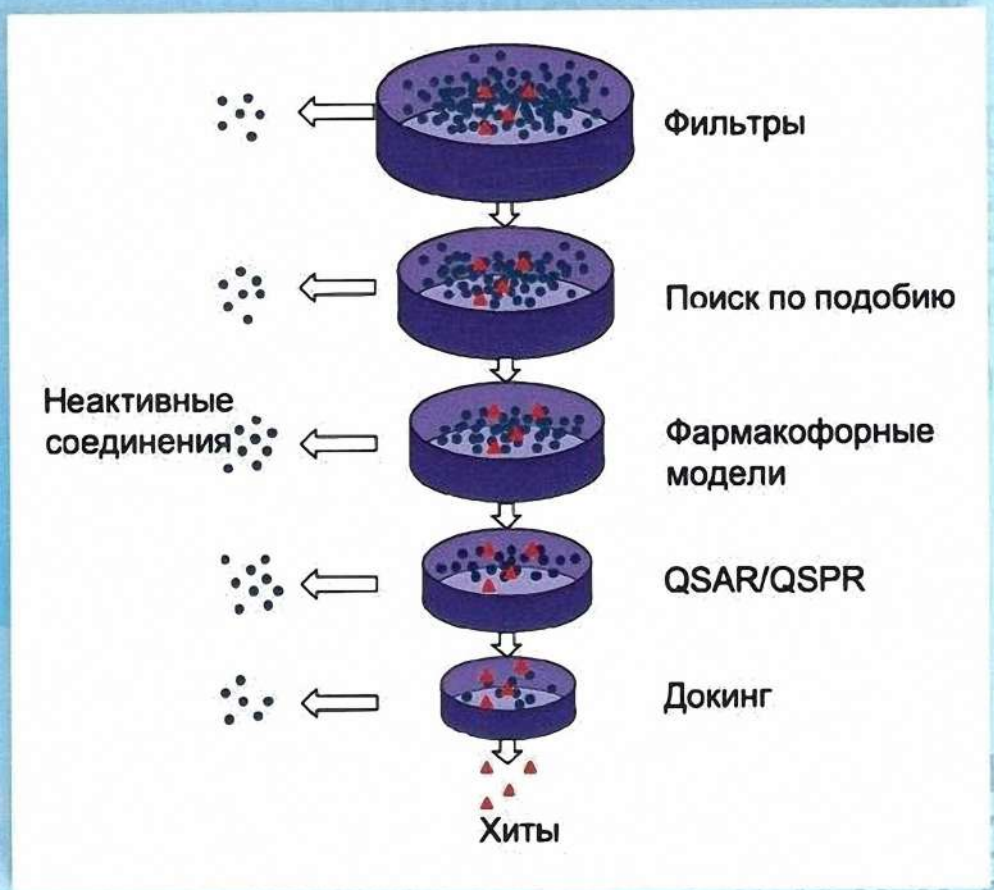
Т.И. МАДЖИДОВ
И.И. БАСКИН
А.А. ВАРНЕК

ХИМИЧЕСКОЕ ПРОСТРАНСТВО И ВИРТУАЛЬНЫЙ СКРИНИНГ

Часть 6

ВВЕДЕНИЕ В ХЕМОИНФОРМАТИКУ



КАЗАНЬ
2020

Т.И. Маджидов, И.И. Баскин, А.А. Варнек

Введение в хемоинформатику

Учебное пособие

Часть 6

Химическое пространство и виртуальный скрининг

Казань
Москва
Страсбург
2020

Версия 1.0

Введение в хемоинформатику: Химическое пространство и виртуальный скрининг: учеб. пособие / Т.И. Маджидов, И.И. Баскин, А.А. Варнек. – Казань, Москва, Страсбург, 2020. – 262 с.

Данное пособие является шестым и заключительным из серии «Введение в хемоинформатику». Оно дает основные сведения о методах анализа и визуализации химического пространства и организации в нем виртуального скрининга химических соединений для поиска новых лекарственных препаратов, материалов, химических веществ с заданными свойствами.

Учебное пособие предназначено для студентов бакалавриата, специалитета и магистратуры, получающих образование в области хемоинформатики и молекулярного моделирования.

Печатная версия пособия:

Введение в хемоинформатику: учеб. пособие / Т.И. Маджидов, И.И. Баскин, А.А. Варнек. – Казань: Изд-во Казан. ун-та, 2019. – Ч. 6: Химическое пространство и виртуальный скрининг. – 238 с.

ПРЕДИСЛОВИЕ

Широкое распространение высокопроизводительного скрининга соединений (HTS, англ. high-throughput screening) и развитие комбинаторной химии позволило существенно ускорить разработку новых лекарственных препаратов. Скрининг позволяет сократить количество тестируемых соединений тысячи раз. В то же время, из тысяч и десятков тысяч прототипов лишь единицы способны удовлетворить множеству требований по активности *in vitro* и *in vivo*, по свойствам ADME (адсорбция, распределение, метаболизм, выделение), селективности действия, отсутствию токсичности и побочных эффектов, а после этого – успешно пройти все фазы клинических испытаний и после этого стать лекарствами.

При разработке новых лекарственных препаратов мечтой химика было бы иметь базу данных, в которой зарегистрированы для всех химических соединений результаты всех видов биологических испытаний. В последние годы, благодаря развитию технологии высокопроизводительного экспериментального скрининга, постановка такой задачи уже не является фантастикой, и при достаточном финансировании ее выполнение вполне осуществимо на крупных фармакологических компаниях при естественных ограничениях на типы биологических испытаний и при использовании доступных библиотек синтезированных химических соединений. К сожалению, опора на библиотеки уже синтезированных соединений является очень существенным ограничителем такого подхода, поскольку не позволяет осуществлять поиск соединений с принципиально новыми структурами, которые бы не были покрыты патентами и имели бы существенные конкурентные преимущества на рынке.

Помимо экспериментального биологического скрининга, существует также *виртуальный скрининг* – процедура полностью компьютеризированного отбора с использованием средств хемо- и биоинформатики молекул, обладающих требуемыми свойствами, из базы существующих или сгенерированных на компьютере химических соединений. Виртуальный скрининг включает в себя использование целого набора вычислительных подходов, направленных на предсказание тех или иных свойств молекул. Сама процедура виртуального скрининга может очень сильно различаться в зависимости от целей, возможностей, объектов и множества других факторов. Хотя первоначально процедура виртуального скрининга использовалась главным образом при разработке новых лекарственных препаратов, в настоящее время она находит широкое применение

совершенно в других областях – при разработке каталитических систем, в дизайне новых материалов, реакций.

Существенным преимуществом виртуального скрининга перед экспериментальным является, кроме большей дешевизны, принципиальная возможность работы с огромным числом ранее не изученных и еще даже не синтезированных соединений, которые могут содержать принципиально новые хемотипы, не имеющие известных структурных аналогов. В то же время, платой за такое преимущество является принципиальная невозможность постановки задачи предсказания всех видов биологической активности для всех синтетически доступных химических соединений из-за астрономического числа последних, сопоставимого с числом частиц во Вселенной. При работе с таким множеством химических соединений на первый план выходят методы хемоинформатики, рассматривающих их как элементы химического пространства.

Данная книга посвящена рассмотрению методов виртуального скрининга как операций в химическом пространстве молекулярных структур. Изложенный в ней материал относится только к подходам, применимым к небольшим молекулам и не требующим знания пространственных структур биологических макромолекул – мишеней для действия низкомолекулярных лигандов. Поэтому методы докинга и моделирования методами молекулярной динамики в книге не рассматриваются.

Книга начинается с рассмотрения объектов химического пространства и отношений сходства между ними. Вторая глава книги посвящена рассмотрению двух типов описания химического пространства – химического пространства графов и химического пространства дескрипторов. При описании химического пространства графов рассматриваются подструктурные (в частности, методы описания химических соединений на основе каркасов и остовов), надструктурные подходы, а также методы передвижения в пространстве с помощью молекулярных мутаций. Также в этом разделе рассмотрены подходы на основе концепции пар соответствия молекул (англ. *molecular matched pairs*). При рассмотрении химического пространства дескрипторов излагаются методы картографии химического пространства на основе карт SOM и GTM, а также обсуждается использование индексов SARI и SALI. Третья глава книги посвящена методам формирования библиотек химических соединений для проведения виртуального скрининга. В частности, рассмотрены разные виды библиотек соединений и способы их формирования, методы генерации химических структур при помощи комбинирования фрагментов (в частности, RECAP и Fragmenter).

Кроме того, рассматривается проблема формирования оптимальных библиотек соединений, а также методы форирования виртуальных комбинаторных библиотек. Глава 4 посвящена фармакофорному анализу. Рассматриваются различные методы определения фармакофоров, как с использованием структур комплексов «белок-лиганд», так и без знания струкур белков. Кроме того, рассмотрен и топологический вариант фармакофоров. Завершающая Глава 5 целиком посвящена различным процедурам виртуального скрининга на основе структур лигандов (англ. ligand-based). В частности, обсуждена концепция виртуального скрининга, введены понятия «воронки» и «фильтров», рассмотрены различные числовые характеристики производительности компонент виртуального скрининга. Приведен список наиболее популярных виртульных библиотек для виртуального скрининга. Далее описаны важнейшие виды фильтров для виртуального скрининга, методы ранжирования молекул с использование 2D-структур (по сходству с активными соединениями, по склонности к обладанию активностью на основе моделей, построенных методами машинного обучения) и 3D-структур молекул (на основе квантового сходства, сходства пространственных форм молекул, сходства молекулярных полей).

Мы хотели бы поблагодарить людей, поддерживавших нашу работу над серией учебных пособий «Введение в хемоинформатику»: заведующего кафедрой органической химии КФУ, член-корр. РАН И.С. Антипина, а также преподавателей магистратуры по хемоинформатике КФУ и Страсбургского университета. Мы очень благодарны людям, непосредственно принимавшим участие в подготовке данной книги: к.х.н. О.В. Климчук за помощь, полезные советы и замечания, аспирантам В.А. Афониной, А.А. Фатыховой, А. Рахимбековой и Р.Н. Мухаметгалиеву, а также Т.И. Сибгатуллиной, студентам Р. Пикалевой и К. Пикалевой за неоценимую помощь в подготовке иллюстративного материала.

Т.И. Маджидов,
И.И. Баскин,
А.А. Варнек

1. ХИМИЧЕСКОЕ ПРОСТРАНСТВО

Хемоинформатика ориентирована на работу с широким кругом объектов, составляющих предмет изучения химической науки: химическими соединениями, их смесями и растворами, химическими реакциями, каталитическими системами, материалами и пр. В дальнейшем мы их будем называть *химическими объектами*. Характерной чертой химических объектов является то, что при заданных внешних условиях каждый из них обладает определенным набором свойств, которые могут быть табулированы в таблицах, либо занесены в электронные базы данных. Примерами таких свойств для химических соединений являются физико-химические свойства (температура кипения, плавления и др.) и разные виды биологический активности, а для химических реакций – константы скорости.

Важнейшими задачами хемоинформатики является прогнозирование свойств химических объектов и конструирование химических объектов, обладающих заданными свойствами. Характерной особенностью хемоинформатики как метода моделирования является то, что прогнозирование свойств объектов производится на основе известных значений свойств других объектов. Принципиальным моментом здесь является то, что для такого переноса свойств необходимо, чтобы между объектами было определено отношение *сходства* (соседства, близости, подобия), которые определяют ту меру, с которой свойства одного объекта могут быть перенесены на другой по принципу: сходные объекты обладают сходными свойствами.

Химическое пространство – это множество химических объектов, между которыми определены отношения сходства (соседства, близости, подобия). Эти отношения придают «структуру» химическому пространству и определяют методы его описания. Именно благодаря ключевому значению, которое играет отношение сходства в прогнозировании свойств химических объектов, понятие химического пространства является центральным для хемоинформатики.

1.1. ОБЪЕКТЫ ХИМИЧЕСКОГО ПРОСТРАНСТВА

Для удобства применения математических методов для представления химических объектов используют различные *математические структуры*. Ключевая роль из них принадлежит

молекулярным графам (основные понятия теории графов в приложении к химическим структурам см. Раздел 2 в Пособии 2), вершины которых соответствуют атомам, а ребра – химическим связям, см. Рис. 1 (а). Метки вершин указывают на тип соответствующего химического элемента, а метки связей характеризуют типы связей. Ключевая роль молекулярных графов обусловлена тремя факторами. Во-первых, они могут быть использованы для однозначной идентификации представителей основного класса химических объектов – химических соединений. Во-вторых, содержащейся в молекулярных графах информации достаточно для получения всех остальных способов представления химических соединений, таких как матрица смежности, строки линейной нотации, молекулярные отпечатки пальцев, фармакофорные графы, вектора дескрипторов, молекулярная геометрия, молекулярные поля и др. В-третьих, они являются основными составными частями для построения практически всех классов хемоинформационных объектов, таких как химические реакции, смеси химических соединений, супрамолекулярные комплексы, сополимеры и др.

Следует отметить, что целый ряд классов хемоинформационных объектов также могут быть представлены при помощи графов, содержащих дополнительные, по сравнению с молекулярными графами, наборы меток для вершин и ребер. Например, супрамолекулярные комплексы могут быть описаны при помощи графов, содержащие специальные «координационные» типы связей, тогда как для кодирования химических реакций могут быть использованы «динамические» связи [1] (см. часть 2.5 пособия 5). Более сложные химические системы, такие как полимеры, сополимеры и смеси, описываются при помощи объектов, включающих в свой состав один или несколько молекулярных графов.

Таким образом, молекулярные графы являются базовым элементом для представления химической информации. Следует, однако, отметить и некоторые недостатки использования графов для этой цели. В частности, с их помощью затруднительно (хотя в ограниченной мере и возможно при помощи определенных эвристических приемов) представление стереохимической информации, а также описание многоцентровых связей. Хотя в качестве универсального метода преодоления такого рода недостатков было предложено использовать аппарат гиперграфов [2], этот подход не получил распространения вследствие большой сложности и трудоемкости.

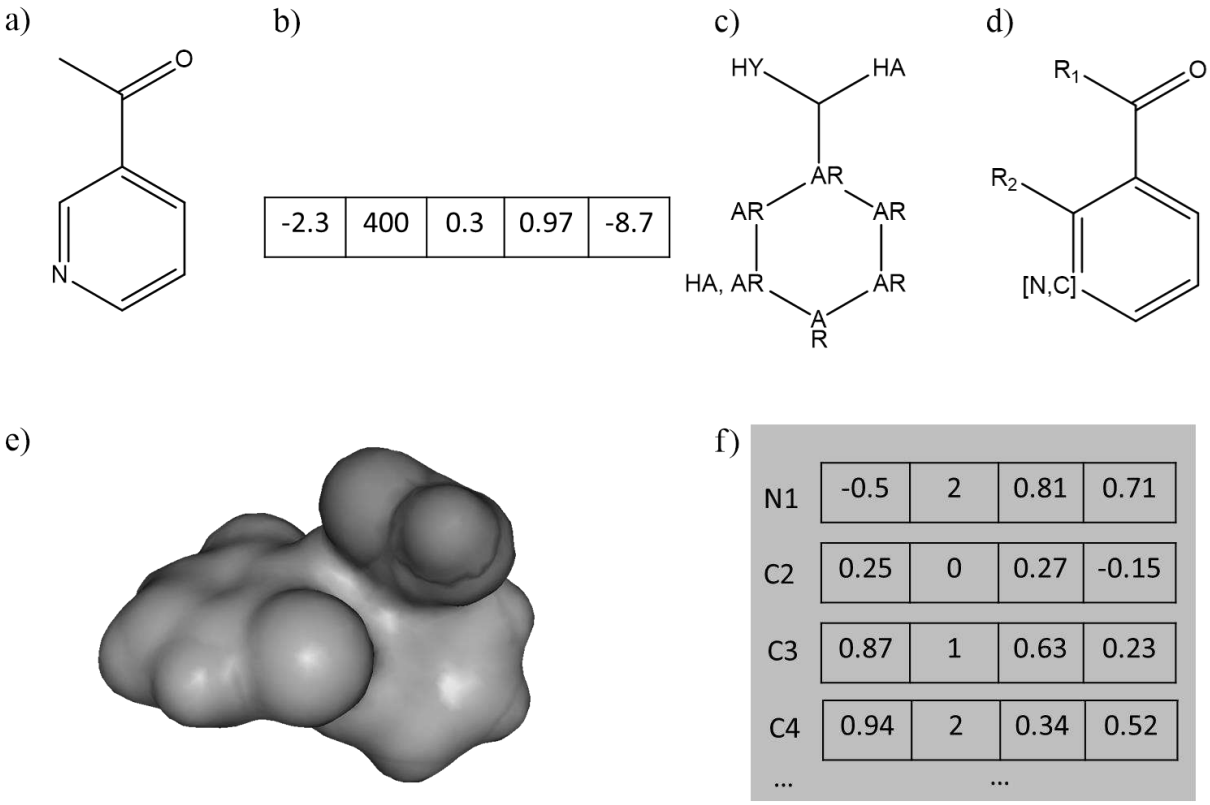


Рис. 1. Математические структуры, используемые для представления химических объектов: a) молекулярный граф, b) вектор дескрипторов, c) фармакофорный граф, в котором центры обозначены как HY-гидрофобный атом, HA-атом – акцептор водородной связи, AR – ароматический атом, d) структура Маркуша, e) изоповерхность электронной плотности, f) локальные дескрипторы атомов.

Если для однозначной идентификации химического (и особенно органического) соединения и хранения его в базе данных наиболее подходит представление при помощи молекулярного графа, то в приложениях, связанных с анализом связи «структура-свойство», для хемоинформационных объектов используют, с целью упрощения применения статистических методов анализа данных, дополнительные математические структуры. В настоящее время основной математической структурой для представления хемоинформационных объектов для этой цели являются вектора дескрипторов [3], см. Рис. 1 (b) (см. раздел 1 пособия 3). Причина очень большой популярности векторов молекулярных дескрипторов в хемоинформатике связана с тем, что: *(а)* вектора чисел, в отличие от графов, легко обрабатывать с помощью стандартных методов хранения и статистического анализа данных с использованием доступных программных средств; (*б*) разработано очень большое число разнообразных типов дескрипторов,

осмысленный выбор из которых позволяет направленно извлекать наиболее релевантную для решения поставленной задачи часть скрытой в молекулярных графах информации; (*в*) вследствие инвариантности значений дескрипторов по отношению к перенумерации вершин графа, их использование избавляет от необходимости решать сложные комбинаторные проблемы, связанные с изоморфизмом графов; (*г*) использование в качестве дескрипторов экспериментальных, либо теоретически предсказанных значений близких, либо зависимых свойств делает модель «структура-свойство» легко интерпретируемой и может приводить к значительному росту ее прогнозирующей способности; (*д*) дескрипторы определяют векторное пространство, с которым значительно легче работать (например, с использованием методов линейной алгебры), чем с дискретным пространством графов. Следует также отметить, что можно подготовить вектора дескрипторов не только для индивидуальных молекул, но и для таких сложных химических систем, как химические реакции [1] и многокомпонентные смеси [4]. Частным случаем векторов дескрипторов являются «молекулярные отпечатки пальцев».

Не следует, однако, забывать и об определенных недостатках использования векторов дескрипторов: (*а*) выбор используемых типов дескрипторов чрезвычайно субъективен и часто диктуется наличием программных средств для их вычисления; (*б*) при неудачном подборе дескрипторов нескольким хемоинформационным объектам могут соответствовать одинаковый набор дескрипторов, что делает их неразличимыми; (*в*) число известных дескрипторов очень велико, и это представляет проблему, даже несмотря на наличие множества методов отбора переменных [5], поскольку всегда существует риск, что отобранные дескрипторы будут нерелевантны, либо избыточны; (*г*) в общем случае нельзя однозначно восстановить молекулярный граф исходя из набора дескрипторов.

Математической структурой, упрощающей применение методов фармакофорного анализа, является фармакофорный граф, вершины которого соответствуют фармакофорным центрам и несут метку, обозначающую их тип (например, H-донор, H-акцептор, катион, анион, алифатическая группа, ароматический цикл и др.), а ребра – топологическому либо геометрическому расстоянию между ними [6], см. Рис. 1 (с). Фармакофорные графы можно использовать как непосредственно при фармакофорном анализе, так и в качестве промежуточной структуры для расчета фармакофорных дескрипторов.

Математической структурой, широко используемой, в частности, в патентном поиске, является структура Маркуша, в которой вершины

могут соответствовать нескольким типам атомов либо целым подструктурам, например, заместителям [7], Рис. 1 (d). То же самое относится и к молекулярным графам, формируемых при построении запросов для поиска в химических базах данных [8].

Математической структурой, упрощающей работу с молекулярными полями, являются гладкие функции от пространственных координат, Рис. 1 (e). Примерами таковых являются одночастичные функции электронной плотности, функция электростатического потенциала, многочисленные молекулярные поля.

Еще одной математической структурой, применяемой в некоторых исследованиях в области хемоинформатики, является множество векторов локальных дескрипторов атомов, Рис. 1 (f). В отличие от векторов дескрипторов, это множество является неупорядоченным, его мощность (количество элементов) зависит от числа атомов в молекуле и поэтому не фиксирована, а в качестве элементов содержит вектора (фиксированной длины) локальных дескрипторов атомов. Такого рода математические структуры иногда называют «*облаками точек*» (англ. point clouds).

1.2. ОТНОШЕНИЯ СХОДСТВА МЕЖДУ ОБЪЕКТАМИ ХИМИЧЕСКОГО ПРОСТРАНСТВА

Молекулярное сходство (или *химическое сходство*) является базовой концепцией хемоинформатики [9, 10]. Она широко используется в виртуальном скрининге и в *in silico* дизайне новых соединений. Такие исследования опираются на принцип: сходные химические объекты имеют сходные свойства [9]. В применении к классификационным проблемам, возникающим в хемоинформатике, это означает, что сходные химические объекты преимущественно относятся к одному классу (т.н. гипотеза компактности классов). Например, сходные по строению химические соединения преимущественно характеризуются одинаковыми типами биологической активности. В приложении к регрессионным проблемам принцип сходства означает, что функция, аппроксимирующая количественную зависимость свойств химических объектов от их строения, должна быть как можно более гладкой. Для количественного описания химического сходства между химическими объектами одного класса вычисляют различные меры сходства, которые могут быть использованы для прогнозирования свойств

химических объектов и организации поиска по сходству в химических базах данных [9].

1.2.1. Уровни отношения сходства

В зависимости от наборов математических условий, налагаемых на меры сходства, можно говорить о трех уровнях отношений сходства: базовый (индексы сходства), метрики и ядра сходства. Каждый из них определяет набор доступных возможностей, причем каждый последующий уровень включает все возможности, предоставляемые предыдущим. Меры сходства будут далее рассмотрены в приложении к различным типам математического описания химических объектов.

1.2.1.1. Базовый уровень отношения сходства

Пусть x и y - химические объекты. На множестве химических объектов можно определить функцию $s(x, y)$, удовлетворяющую для любых x и y следующим условиям: (i) $0 <= s(x, y) <= 1$, т.е. значение функции сходства должно лежать в интервале от 0 до 1; (ii) $s(x, x) = 1$, т.е. значение функции равно единице для одинаковых объектов, (iii) $s(x, y) \rightarrow 0$ для сильно различающихся объектов x и y. В этом случае будем говорить, что функция $s(x, y)$ определяет соотношение сходства между объектами и полученное число называется *индексом сходства*.

Функции сходства удобно использовать для количественного описания известных на качественном уровне закономерностей. Например, если известно, что обладание определенным видом биологической активности определяется пространственной формой молекул, то используемая в этом случае функция сходства может, например, сравнивать формы молекул и давать числовую оценку степени их близости в интервале от 0 до 1, достигая 1 только в случае идеального совпадения форм. Как только сконструирована такая функция подобия, ее можно использовать для виртуального скрининга баз химических соединений путем упорядочивания их по уменьшению ее значения. Эта же функция может быть использована для прогнозирования свойств соединений по методу ближайших соседей, а также для проведения иерархического кластерного анализа баз данных (см. описание процедуры моделирования в разделе 2 пособия 3, а также описание методов обучения в пособии 4). Полученными в последнем случае дендрограммами можно пользоваться для визуализации химического пространства (см. раздел 2.14.1.1 пособия

4). Для этой же цели можно воспользоваться т. н. неметрическими методами нелинейного шкалирования (см. раздел 2.14.2.2 пособия 4).

Следует, однако, подчеркнуть, что наличие лишь базового уровня отношения сходства дает возможность воспользоваться лишь небольшой долей возможностей, предоставляемых современными методами статистического анализа данных и машинного обучения. Кроме того, одного базового уровня недостаточно для применения эффективных алгоритмов поиска по подобию в базах данных. Именно поэтому очень важно пользоваться мерами сходства как можно более высокого уровня.

1.2.1.2. Метрика как отношение сходства

По определению, *метрика* – это функция, определенная для пары элементов множества и задающее расстояние между ними. Для любых x, y, z эта функция должна удовлетворять следующим условиям: (i) $d(x, y) = 0$ только при одинаковых x и y (аксиома тождества); (ii) $d(x, y) = d(y, x)$ (аксиома симметрии); (iii) $d(x, z) \leq d(x, y) + d(y, z)$ (неравенство треугольника, называемое также правилом треугольника). Из этих аксиом также вытекает свойство неотрицательности расстояния: $d(x, y) \geq 0$. Химическое пространство, на котором задана функция расстоянии (метрика), называется *метрическим химическим пространством*.

Функция расстояния в химическом пространстве определяет меру удаленности химических объектов друг от друга. В том случае, если на множестве химических объектов задана метрика $d(x, y)$, принимающая максимальное значение 1, то функция сходства $s(x, y)$ может быть определена следующим образом: $s(x, y) = 1 - d(x, y)$. В этом случае функция сходства называется дополнением метрики. Аналогично, имея функцию сходства, принимающую значение 1 для идентичных объектов, можно определить функцию расстояния как $d(x, y) = 1 - s(x, y)$. Условием того, чтобы значения определенной таким образом функции $d(x, y)$ могли считаться расстояниями в метрическом химическом пространстве, является выполнение вышеперечисленных аксиом (i) - (iii).

Наличие свойства метрики у функций расстояния между химическими объектами дает возможность считать сходные объекты близко расположенными друг к другу в метрическом химическом пространстве, что открывает множество дополнительных возможностей по сравнению с применением функций сходства базового уровня.

Прежде всего, наличие свойств метрики у функции расстояния позволяет эффективно организовать поиск по подобию в больших базах данных (см. раздел 1.2.5 пособия 2). Если имеется в наличии лишь отношение подобия базового уровня, то для поиска наиболее схожих с запросом объектов базы данных необходимо рассчитать меру подобия запроса с каждым объектом базы данных, ранжировать все объекты по найденным значениям индекса сходства и лишь после этого отобрать несколько объектов с наибольшими значениями меры подобия. Необходимый для обработки подобного запроса объем вычислений пропорционален количеству объектов в базе данных. В случае больших баз данных, содержащих многие миллионы объектов, это может привести к огромным вычислительным затратам, делающих подобных поиск неосуществимым на практике.

Если же функция расстояния является метрикой, то можно воспользоваться свойствами метрического пространства для быстрого поиска наиболее близко расположенных к шаблону объектов. Ускорение поиска осуществляется благодаря эффективному использованию неравенства треугольника. Действительно, если нас интересуют ближайшие к объекту x соседи по химическому пространству, и нам известно, что объект z расположен далеко от x, но близко к y, то, вследствие неравенства треугольника $d(x, y) \geq d(x, z) - d(y, z)$, y не может быть расположен близко к x, и, следовательно, нет необходимости вычислять расстояние между x и y. Пользуясь подобным приемом, можно значительно сократить объем вычислений. Наиболее известный алгоритм для осуществления этого основан на использовании т.н. k-мерных деревьев [11] для организации доступа к ближайшим объектам за время, пропорциональное логарифму их общего числа в базе данных. Это дает возможность эффективно осуществлять поиск по сходству даже для баз данных очень большого размера. С помощью этого же приема можно значительно ускорить прогнозирование свойств объектов по методу k ближайших соседей.

Наличие свойства метрики у отношения сходства открывает также возможность строить карты химического пространства на плоскости либо в 3D-объеме при помощи целого набора линейных и нелинейных методов: метода многомерного шкалирования (англ. *Multi-Dimensional Scaling - MDS*), картирования по Сэммону (англ. *Sammon Mapping - SM*), стохастического встраивания соседей (англ. *Stochastic Neighbor Embedding - SNE*) и др. Сочетание метрических свойств химического пространства с принципом «сходные химические объекты обладают сходными свойствами» приводит к появлению на полученных картах зон, в которых преобладают химические

соединения с одинаковым типом биологической активности. Это позволяет строить карты химического пространства, напоминающие политические карты мира, в которых области с преобладанием определенного вида биологической активности окрашены определенным цветом. Такие карты могут быть использованы при виртуальном скрининге больших электронных библиотек химических соединений.

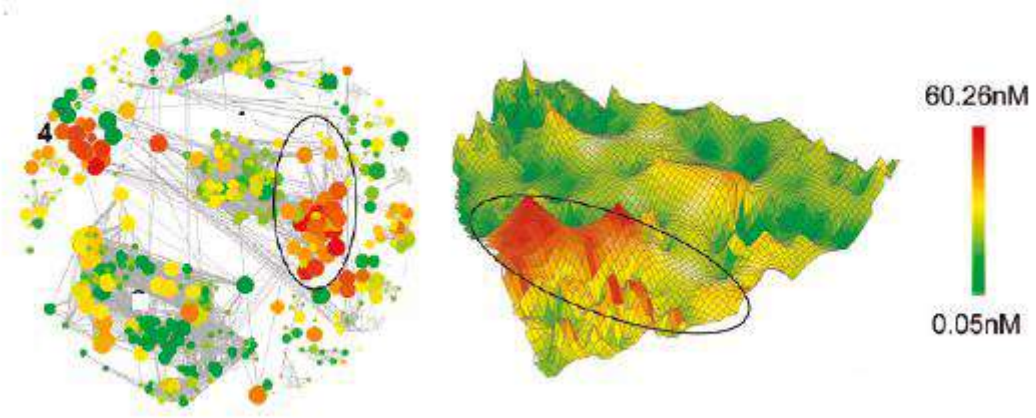


Рис. 2. Пример карты, построенной с использованием графа сходства в работе [12]. Градация цвета характеризует активность соединений. Рисунок из публикации приводится с разрешения издательства. Copyright (2011) American Chemical Society.

В качестве примера, в котором использовался данный подход, можно привести работу [12]. В ней проводился анализ активности аденозиновых рецепторов с использованием графа сходства, Рис. 2. Пример карты, построенной с использованием графа сходства в работе [12].Рис. 2. Для построения последнего определялось сходство молекул с использованием индекса Танимото, являющегося дополнением метрики – расстояния Сергела. В качестве дескрипторов брались структурные ключи MACCS [13]. Каждая молекула представлялась на карте вершиной. Вершины были связаны ребром, если сходство объектов превышала 0.8. Положение объектов на карте далее оптимизировалось так, чтобы тесно связанные группы образовывали компактные кластера. Как видно на рисунке, образующиеся кластеры, как того и требует принцип сходства, содержат объекты, которые имеют близкие значения активности (обозначены цветом на Рис. 2 слева).

В том случае, когда биологическая активность выражена количественно, например, при помощи величин EC_{50} и IC_{50}, метрические свойства химического пространства позволяют с

помощью вышеупомянутых подходов, в дополнение к отображению множества объектов на двухмерную карту, строить карты ландшафтов, в которых величина изучаемого свойства закодирована третьей координатой – высотой, Рис. 2 справа. В этом случае зависимость свойства от структуры объекта представлена «геодезической» картой «горной местности». Использование подобных карт помогает вести направленную оптимизацию свойств химических объектов путем «взбирания на вершины гор». Подобные карты могут быть также построены и для случая качественного анализа биологической активности. В последнем случае «высота» определяется оценкой вероятности принадлежности к определенному классу, определяемому типом биологической активности, либо численным значением активности (IC_{50} или других).

Следует подчеркнуть, что вышеупомянутые способы построения карт никоим образом не ограничены областью медицинской химии, а могут быть применены к любым классам химических объектов.

Еще одним из преимуществ, связанных с наличием метрических свойств у химического пространства, является возможность локализации областей химического пространства, в которых содержатся резкие всплески или падения активности соответствующих соединений (называемые пиками и провалами активности, англ. *activity cliffs*). Последние являются проявлениями нарушений принципа сходства и более подробно описываются далее.

1.2.1.3. Ядро как отношение сходства

Ядро $k(x,x')$ (называемое также ядром Мерсера) - это симметричная функция (т.е. $k(x,x') = k(x',x)$), показывающая сходство объектов x и x', для которой для любого набора объектов x_i матрица Грама **K**:

$$\mathbf{K} = \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_N) \\ \vdots & \ddots & \vdots \\ k(x_N,x_1) & \cdots & k(x_N,x_N) \end{pmatrix} \qquad (1)$$

является полу-положительно определенной (англ. *semi-positive definite*), что означает, что у матрицы **K** не может быть ни одного отрицательного собственного значения. В том случае, если все объекты разные, матрица **K** должна быть положительно определена, т.е. все собственные ее значения должны быть положительны. Следует, однако, иметь в виду, что отсутствие отрицательных собственных значений у полученной для одного набора объектов матрицы **K** является необходимым, но отнюдь не достаточным условием того, что

функция сходства является ядром. Если хоть для одного набора объектов эта матрица содержит хотя бы одно отрицательное собственное число, то это означает, что соответствующая функция сходства ядром не является. Следовательно, даже если у матрицы **K** для одного набора данных отсутствуют отрицательные собственные числа, то это не означает, что и для всех других наборов данных они будут отсутствовать. Таким образом, показать, является ли мера сходства ядром, можно только при помощи строгого математического доказательства. Ядро редко используется как способ оценки сходства между молекулами в виртуальном скрининге или в базах данных – для них они не имеют существенного преимущества перед другими способами оценки сходства, например, метриками. Основное применение ядро сходства находит в машинном обучении в так называемых ядерных методах машинного обучения (см. раздел 2.5 в пособии 4). Ядро сходства позволяет переходить от исходного дескрипторного пространства к новым, более высокоразмерным пространствам (так называемым «спрямляющим», или «пространствам признаков»), в которых можно провести линейное разделение при решении классификационных задач или линейную регрессию, когда этого нельзя сделать в исходном пространстве (Рис. 3).

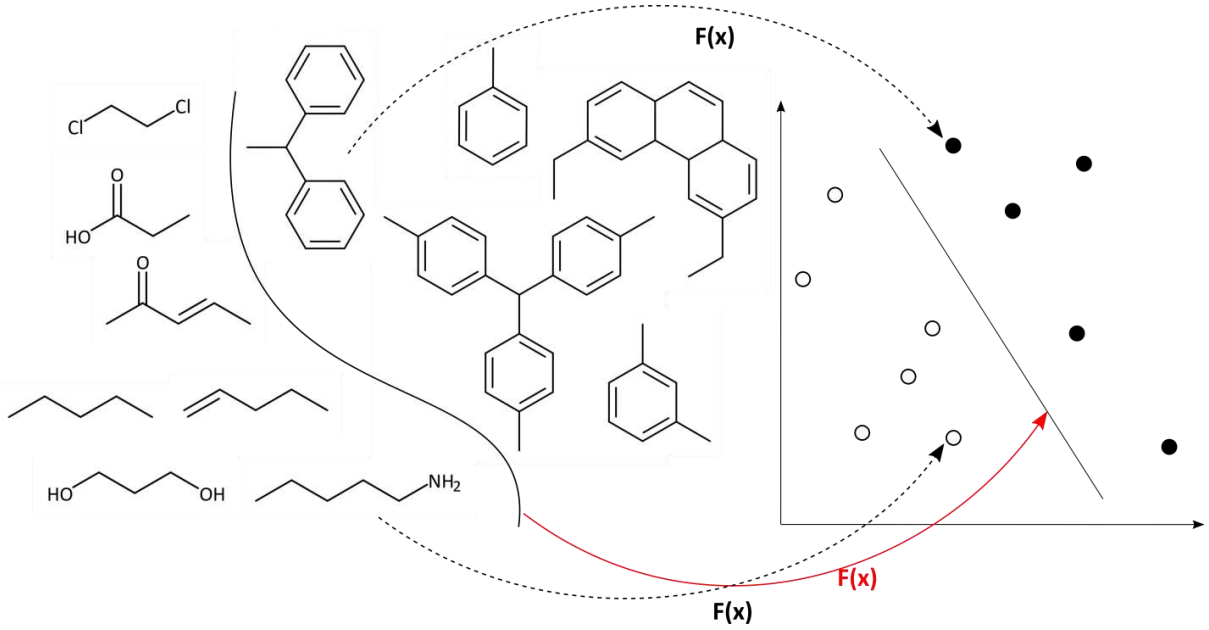


Рис. 3. Отображение химических объектов в пространство признаков.

Покажем, как ядра позволяют переходить в высокоразмерные пространства. Предположим, мы имеем D-мерное дескрипторное пространство, то есть каждый объект i представлен вектором признаков (дескрипторов) $\boldsymbol{x}_i$. Скалярное произведение векторов для двух молекул находится как:

$$\langle \boldsymbol{x}_i | \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^T \boldsymbol{x}_j \quad (2)$$

На основе исходного дескрипторного пространства можно построить бесконечное количество новых пространств, например, исходный набор дескрипторов для объекта $\boldsymbol{x}_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,D}\}$ заменяется всеми их попарными произведениями. Можно считать, что мы задали функцию отображения $\Phi(\boldsymbol{x})$, которая переводит исходное дескрипторное пространство в новое. Полученное пространство будет содержать гораздо большее количество дескрипторов:

$$\Phi(\boldsymbol{x}_i) = \{ x_{i,1}x_{i,1}, x_{i,1}x_{i,2}, \dots, x_{i,1}x_{i,D}, x_{i,2}x_{i,1}, x_{i,2}x_{i,2}, \dots, x_{i,D}x_{i,D} \} \quad (3)$$

Преимущество такого подхода состоит в том, что если в исходном пространстве была квадратичная зависимость между каким-либо свойством и одним из дескрипторов, то в новом пространстве эта зависимость уже будет линейной. В нем уже оказывается возможным разделить плоскостью объекты, которые в исходном пространстве можно было разделить только окружностью (см. Рис. 4).

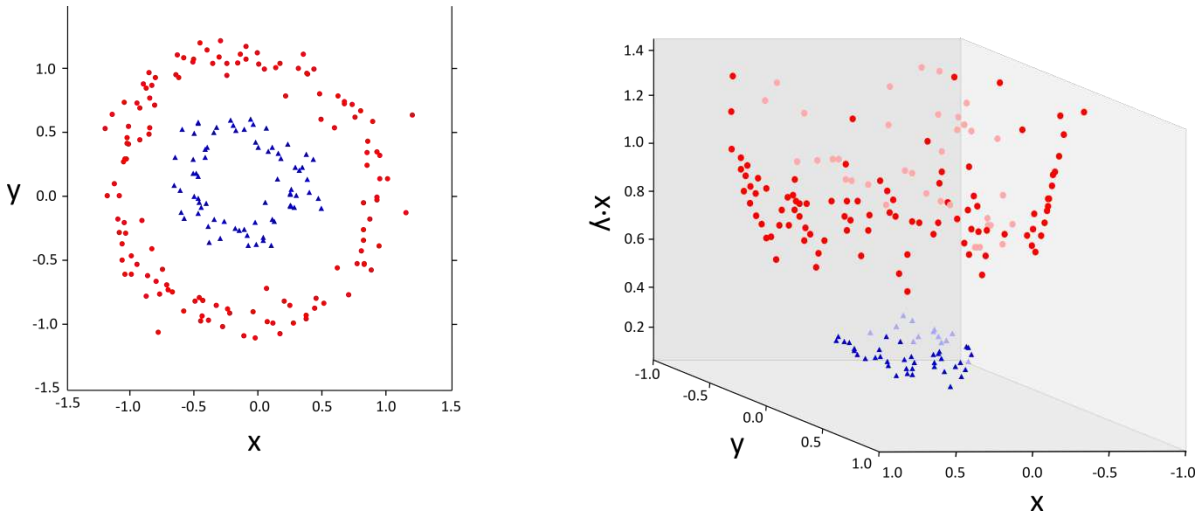


Рис. 4. Слева: объекты двух классов (кружки или треугольники) нельзя разделить прямой или плоскостью. Справа: если добавить третий дескриптор, равный произведению двух исходных дескрипторов (приведен по вертикальной оси), то разделение объектов плоскостью становится возможным.

Скалярное произведение векторов в этом новом пространстве и является ядром сходства:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \langle \Phi(\boldsymbol{x}_i) | \Phi(\boldsymbol{x}_j) \rangle = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j) \quad (4)$$

Если отображение $\Phi(\boldsymbol{x})$ оставляет исходное пространство неизменным, то ядром выступает скалярное произведение в исходном

пространстве дескрипторов. Это ядро называют линейным. Таким образом, если метод машинного обучения требует расчета скалярного произведения векторов, как, например, метод опорных векторов (см. раздел 2.5 пособия 4) или гауссовых процессов для восстановления регрессии (см. раздел 2.6.3 пособия 4), то вместо скалярного произведения можно подставить другое ядро сходства. Это позволит искать нелинейные зависимости в исходном пространстве.

Описанный выше способ определения положения объектов в новом пространстве путем вычисления в явном виде всех их координат $\Phi(\boldsymbol{x})$, например, по формуле (3), с последующим вычислением их скалярных произведений является крайне неэффективным в вычислительном плане из-за большой размерности таких пространств. Более того, для некоторых типов ядер такие вычисления являются принципиально невозможными из-за бесконечной размерности соответствующих пространств. К счастью, существует очень эффективный непрямой способ вычисления скалярных произведений в таких высокоразмерных пространствах, не требующий предварительного вычисления всех координат $\Phi(\boldsymbol{x})$. Рассмотрим этот способ на конкретном примере т.н. квадратичного ядра. Предположим, что имеются два вектора $\boldsymbol{x}$ и $\boldsymbol{z}$, тогда вычислим и упростим следующее выражение:

$$\langle \boldsymbol{x}|\boldsymbol{z}\rangle^2 = (\boldsymbol{x}^T\boldsymbol{z})^2 = \boldsymbol{x}^T\boldsymbol{z}\boldsymbol{x}^T\boldsymbol{z} = \sum_i^D x_i z_i \cdot \sum_j^D x_j z_j = \sum_{i,j} (x_i x_j)(z_i z_j) = \langle \varphi(\boldsymbol{x})|\varphi(\boldsymbol{z})\rangle \quad (5)$$

Заметим, что в конце мы получили скалярное произведение в новом пространстве, в котором вместо исходных дескрипторов берутся все попарные перемножения. Таким образом, чтобы вычислить скалярное произведение в рассмотренном выше пространстве, состоящем из попарных произведений дескрипторов, достаточно возвести в квадрат скалярное произведение векторов в исходном пространстве! Более того, аналогично можно показать, что при возведении скалярного произведения в куб, получится ядро, соответствующее скалярному произведению во многомерном пространстве, содержащем в качестве координат произведения всех троек и т. д. Такое ядро называется полиномиальным. Подобный прием получил название «трюк с ядрами» (англ. kernel trick) и в настоящее время активно используется при анализе данных в рамках ядерных методов машинного обучения. Этот результат, несмотря на

кажущуюся странность, является не частным случаем, а общим свойством любых ядер.

Согласно теореме Мерсера, для любого ядра $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ может быть определено такое пространство, в которое объекты $\boldsymbol{x}$ исходного пространства могут быть отображены при помощи функции отображения $\Phi(\boldsymbol{x})$ таким образом, чтобы скалярное произведение образов объектов в нем было численно равно значению функции ядра для них:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i) | \Phi(\boldsymbol{x}_j) \rangle \qquad (6)$$

На Рис. 3 схематически показано отображение химических объектов исходного пространства (например, пространства молекулярных графов, описывающих химические структуры) в линейное пространство функций, в котором определена операция скалярного произведения векторов (подобные пространства принято называть пространствами Гильберта в честь известного математика Давида Гильберта, впервые их исследовавшего). В математике пространства, задаваемые функциями ядер, называют *представляющими ядра Гильбертовыми пространствами* (англ. *RKHS* - Representing Kernels Hilbert Spaces), однако в большинстве прикладных публикаций их упрощенно называют *пространствами признаков* (англ. feature spaces). В отечественной литературе эти пространства также называют *спрямляющими*, поскольку они являются линейными, подчиняющимися законам линейной алгебры, и потому любая сколь угодно сложная нелинейная зависимость между описывающими объекты переменными в исходном пространстве превращается в линейную зависимость (т.е. спрямляется) в некотором пространстве признаков.

Фундаментальное значение понятия ядра как меры сходства в хемоинформатике заключается в том, что именно оно определяет геометрию химического пространства. Хотя метрика вводит понятие расстояния между объектами, но этого еще недостаточно для описания их взаимного расположения в пространстве. Ядра, находясь на более высоком уровне отношений сходства, задают, наряду с расстояниями, еще и углы, что позволяет задать для химического пространства полноценную геометрию. Рассмотрим далее, каким образом это может быть осуществлено.

Как показано выше, ядра задают в неявном виде отображение химических объектов в пространство признаков, в котором определена операция скалярного умножения векторов. Благодаря такому отображению оказывается возможным работать с химическими

объектами как с обычными векторами и, следовательно, осуществлять любые геометрические построения в химическом пространстве. В частности, расстояние в химическом пространстве может быть определено по формуле:

$$d(x,y) = \sqrt{< \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) >_{H_k}} = \sqrt{k(x,x) + k(y,y) - 2k(x,y)} \qquad (7)$$

Определяемое таким образом расстояние между химическими объектами всегда обладает свойствами метрики, что в общем случае следует из метрических свойств Гильбертова пространства. Угол между объектами может быть определен по формуле:

$$\alpha(x,y,z) = \arccos \frac{< \Phi(x) - \Phi(y), \Phi(z) - \Phi(y) >_{H_k}}{d(x,y) \cdot d(y,z)} = \\ = \arccos \frac{k(x,z) + k(y,y) - k(x,y) - k(y,z)}{d(x,y) \cdot d(y,z)} \qquad (8)$$

При помощи ядер можно также задавать гиперплоскости и определять положение других объектов относительно их, что находит широкое применение при построении классификационных моделей для химических объектов, например, при помощи метода опорных векторов. Таким образом, ядра позволяют определить аналитическую геометрию в химическом пространстве. Эта геометрия, конечно, не столь богата, как в 3-мерном евклидовом пространстве. В частности, в ней нельзя определить векторное произведение для произвольного числа измерений. Тем не менее, для решения задач химической картографии и построения моделей «структура-свойство» ее вполне достаточно.

1.2.2. Отношения сходства для различных математических представлений химических объектов

Рассмотренные выше примеры, демонстрирующие уровни сходства между объектами, исходили из того, что объекты были представлены набором дескрипторов. Несмотря на то, что отношения сходства чаще всего действительно находятся на основе дескрипторного представления молекул, этот способ является далеко не единственным. Ниже мы рассмотрим, каким образом можно вычислять отношения сходства между объектами, представленными с помощью различных математических структур: графов, векторов дескрипторов, функций.

1.2.2.1. Отношения сходства для представления химических объектов в виде графов

Как уже отмечалось, графы являются базовым способом представления большинства химических объектов. Оценка отношений сходства между графами основывается на простой логике: два графа тем более подобны, чем они больше имеют общих фрагментов и чем эти общие фрагменты больше по размеру (то есть содержат больше вершин). С использованием этого могут быть найдены коэффициенты сходства для двух графов, в том числе метрики, а также ядра сходства. Последние позволяют строить модели с использованием ядерных методов машинного обучения, не вычисляя для химических объектов вектора дескрипторов фиксированного размера.

Отношения сходства для графов на основе максимальных общих подграфов

Сходство между объектами, представленными в виде графов (в первую очередь, молекулами), может быть оценено путем определения максимального общего подграфа (англ. *Maximum Common Substructure, MCS*) для них. Максимальный общий подграф двух графов *U* и *H* – это такой граф, который является подграфом в *U* и *H* и при этом содержит максимально возможное число вершин (Рис. 5).

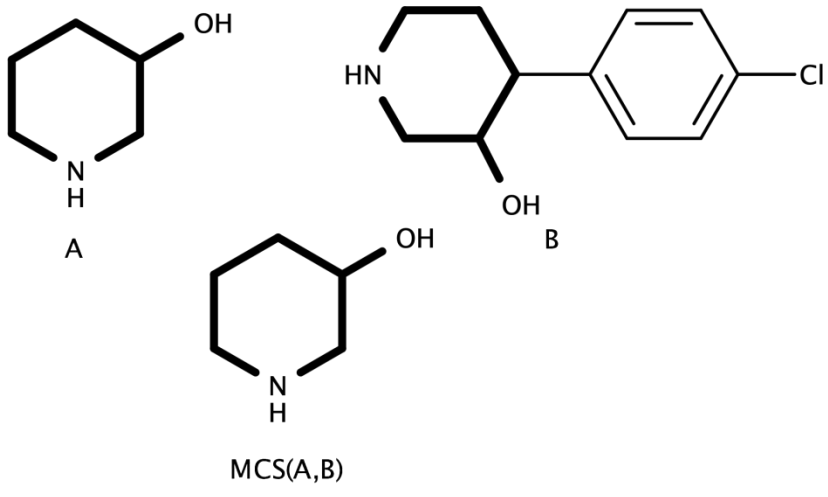


Рис. 5. Максимальный подграф MCS(A,B) для двух молекул A и B.

Оценка сходства молекулярных графов путем выявления максимального общего подграфа для них является одним из самых часто используемых способов расчета мер сходства с использованием графового представления молекул. В этом случае вычисление индекса

сходства молекулярных графов или расстояния между ними ведется на основе сравнения мощности (числа вершин, ребер или их суммы) максимальной общей подструктуры и исходных графов:

- индекс Танимото:

$$S_T(A,B) = \frac{|MCS(A,B)|}{|A| + |B| - |MCS(A,B)|} \quad (9)$$

- индекс Шимкевича-Симпсона:

$$S_{SS}(A,B) = \frac{|MCS(A,B)|}{\min(|A|,|B|)} \quad (10)$$

- индекс косинуса (Карбо):

$$S_C(A,B) = \frac{|MCS(A,B)|}{\sqrt{|A| \cdot |B|}} \quad (11)$$

где $|MCS(A,B)|$, $|A|$ и $|B|$ – мощность максимальной общей подструктуры и исходных графов. Как видно из представленных формул, значения индексов сходства лежат в диапазоне от 0 до 1. Следует также отметить, что в общем случае максимальных общих подструктур для одной и той же пары графов может быть несколько, что, однако, не влияет на значение вышеприведенных индексов, поскольку мощности всех таких подструктур одинаковы.

Вычислительная сложность алгоритма выявления максимальной общей подструктуры сильно ограничивает применение этих индексов сходства, однако они используются в отдельных публикациях для поиска по подобию [14] и кластеризации соединений в базах данных [15].

Известно несколько алгоритмов поиска максимальной общей подструктуры, как точные, так и приблизительные [16]. Точные алгоритмы являются в конечном итоге комбинаторными, их сложность быстро растет с числом вершин в графе, так что для больших графов время, необходимое на выявление максимальной общей подструктуры, может быть непрактично большим. С другой стороны, приблизительные методы, обладающие потенциально большей скоростью работы, не гарантируют, что решение точное, т.е. полученный граф на самом деле содержит максимально возможное число вершин и ребер.

Наиболее эффективный и наиболее распространенный алгоритм точного выявления максимальной общей подструктуры основан на поиске максимальной клики. Основная идея алгоритма заключается в том, что на первом этапе можно найти соответствующие друг другу ребра и вершины графа (то есть допустимые отображения графов друг на друга) с использованием графа совместимости (см. ниже «тензорное

произведение графов»), с последующим поиском максимальной клики с помощью, как правило, алгоритма Брона-Кирбоша [17]. Алгоритм поиска максимальной общей подструктуры описан в разделе 2.3.3 пособия 2.

Графовые ядра

Графовые ядра представляют собой функции $k(G, H)$ для оценки сходства между парой объектов с использованием представлений химических объектов в виде графов G и H. Как было отмечено выше, эти функции должны удовлетворять теореме Мерсера, и поэтому они могут быть представлены как скалярное произведение в некотором пространстве [18]. Основное применение эти методы находят в моделировании «структура-свойство»: например, в работе [19] они использовались для предсказания констант диссоциации. Отказ от необходимости использования стандартного подхода к представлению молекул с помощью вектора дескрипторов фиксированного размера, обладающего определенными недостатками (см. раздел 3.1.1 пособия 4), является основной мотивацией использования графовых ядер, что привело к разработке множества ядер такого типа, не все из которых, однако, нашли применение в хемоинформатике. В то же время расчет графовых ядер, в особенности для больших баз данных, происходит существенно медленнее расчета ядер с использованием дескрипторов, что является основным ограничивающим фактором для их широкого внедрения.

Суть расчета многих графовых ядер сводится к подсчету встречаемости фрагментов определенной топологии (цепочек, деревьев, циклов и т.п.), что находит отражение в названии ядра. Особенность их расчета заключается в том, что фрагменты эти явным образом не создаются, то есть задача не сводится к расчету фрагментных дескрипторов. Можно выделить следующие типы графовых ядер (список не полон, более подробный обзор можно найти в статье [20]):

- *графовые ядра на случайных маршрутах* (англ. random walk graph kernels), которые подсчитывают число общих маршрутов определенной длины. Среди графовых ядер такого типа наибольшее распространение получили маргинализованные (англ. marginalized) графовые ядра, подсчитывающие число общих маршрутов любой длины для двух графов, причем вес маршрутов большой длины снижается для сходимости суммы. Ниже мы рассмотрим этот тип ядер.

- *графовые ядра на графах-деревьях* (англ. tree-based graph kernels или tree pattern kernels). Как следует из названия, их вычисление требует подсчета числа общих подграфов-деревьев двух графов.

- *графовые ядра на циклических фрагментах* (англ. cyclic pattern graph kernels), которые разбивают графы на набор подграфов-циклов и деревьев с последующим подсчетом общих фрагментов.

- *графовые ядра оптимального назначения* (англ. optimal assignment graph kernels) [19], для вычисления которых сначала оценивается сходство вершин, а далее решается задача об оптимальных назначениях вершин, то есть определяются вершины двух графов, наиболее точно соответствующие друг другу, и далее величины, описывающие сходство этих вершин, суммируются (или усредняются).

Рассмотрим в качестве примера маргинализованные графовые ядра на случайных маршрутах, которые являются одними из наиболее часто встречающимися в публикациях. В хемоинформатике они использовались для предсказания мутагенности [21], причем модели с их участием, в целом, показали качество предсказания не хуже, чем при использовании дескрипторного описания молекул. В то же время, в работе [22] при моделировании профилей биологической активности использование графовых ядер не продемонстрирована каких-либо преимуществ перед альтернативными подходами. Алгоритм их вычисления основан на создании так называемого *тензорного произведения графов* (англ. tensor product of graphs, встречается также direct product, modular product). Предположим, имеются два графа U и H, содержащие набор из $V(U)$ и $V(H)$ вершин (обозначим их (u_i) и (v_i)), а также набор ребер (e_i^U) и (e_i^H) соответственно. Каждое ребро k в каждом из этих графов задается как пара смежных ребру вершин i и j, то есть $e_k^U = (u_i, u_j)$, $e_k^H = (v_i, v_j)$. Тензорное произведение графов, обозначаемое обычно как $U \times H$ – это новый граф $T=U \times H$, который содержит вершины, соответствующие всем парам вершин исходных графов, то есть $V(T) = V(U) \cdot V(H)$, причем каждая вершина графа тензорного произведения v_k^T определяется как пара вершин исходных графов (u_i, v_j). Две вершины графа T (u_i, v_j) и (u_k, v_l) связаны ребром, если соответствующие пары вершин исходных графов (u_i, u_k) и (v_j, v_k) тоже связаны ребром. В графе произведения T логично оставлять только вершины для пар атомов, соответствующих одному химическому элементу, а ребра – если метки ребер (типы связей) в исходных графах совпадают. В этом случае граф тензорного

произведения соответствует графу совместимости, рассмотренному в разделе 2.3.3 пособия 2.

Использование тензорного произведения графов удобно потому, что любой маршрут на этом графе одновременно соответствует маршрутам на исходных графах, Рис. 6. Чем больше общих маршрутов содержат два графа, то есть чем больше маршрутов можно провести на графе тензорного произведения, тем более эти графы сходны. Таким образом, вычисление ядра сводится к подсчету числа маршрутов в графе *T*.

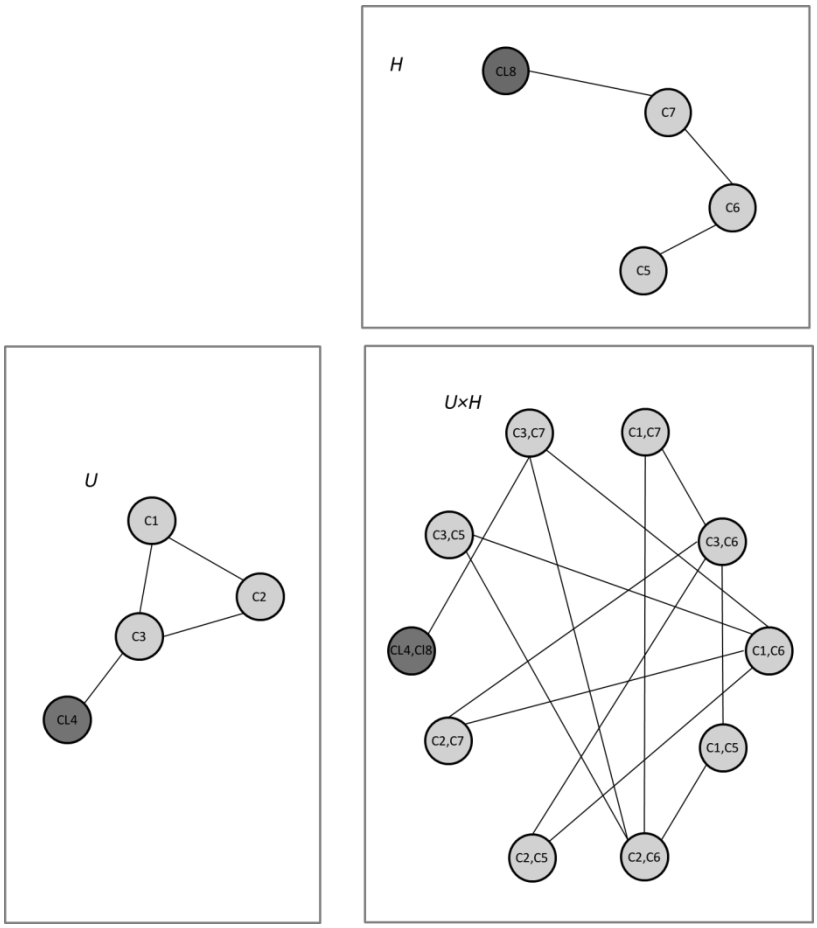


Рис. 6. Графы пропилхлорида и циклопропилхлорида и их тензорное произведение. Для удобства понимания номера атомов в исходных молекулах не пересекаются. Следует обратить внимание, что маршрут (Cl:4,Cl:8)-(C:3,C:7)-(C:1,C:6) соответствует маршруту Cl4-C3-C1 в графе *U* и Cl:8-C:7-C:6 в графе *H*.

Далее для простоты будем полагать, что вершины и ребра графов не имеют меток. В этом случае матрица смежности графа *T* тензорного произведения графов *U* и *H* является кронекеровским произведением

матриц смежностей исходных графов[1], $W = A(U) \otimes A(H)$. Для подсчета числа общих маршрутов удобно использовать следующее свойство матрицы смежности: будучи возведенной в степень n, она дает новую матрицу W^{n}[2], в которой элемент $[W^n]_{ij}$ равен числу маршрутов, соединяющих вершину i и j. Саму же матрицу смежности можно рассматривать как матрицу W^1, описывающую маршруты длиной 1. Таким образом, чтобы найти число маршрутов длины n в графе T, необходимо просуммировать элементы матрицы W^n, или, что эквивалентно, умножить матрицу слева и справа на вектор, состоящий из единиц (обозначим его как $\mathbf{1}$):

$$k^n(U,H) = \sum_{i,j}^{V(T)} W^n = \mathbf{1}^T W^n \mathbf{1} \qquad (12)$$

Поскольку число маршрутов велико, для удобства можно поделить полученное число на квадрат числа вершин матрицы, то есть $V(T)^2 = V(U)^2 V(H)^2$:

$$k^n(U,H) = \frac{1}{V(U)^2 V(H)^2} \mathbf{1}^T W^n \mathbf{1} \qquad (13)$$

Альтернативой является деление каждой строки матрицы смежности на степень вершины (что можно также представить операцией $\widetilde{W} = D^{-1}W$, где D – диагональная матрица степеней вершин). Полученная таким образом «масштабированная» матрица смежности $\widetilde{W}$ отражает вероятность перехода из одной вершины в другую в ходе случайного процесса, являющегося цепью Маркова [3], в котором матрица $\widetilde{W}$ представляет собой матрицу вероятностей перехода.

Как было уже отмечено, графовым ядром можно считать общее число общих маршрутов любой длины – от 1 до ∞ – на двух графах:

---

[1] Произведение Кронекера двух матриц – это блочная матрица, содержащая произведение элемента одной матрицы на другую матрицу. Подробнее с произведением Кронекера можно ознакомиться в Википедии: https://ru.wikipedia.org/wiki/Произведение_Кронекера

[2] Возведение матрицы A в степень n есть продукт умножения матрицы A на саму себя n раз.

[3] Цепь Маркова – это последовательность случайных событий с конечным или счётным числом исходов, характеризующаяся тем свойством, что при фиксированном текущем состоянии следующее состояние независимо от предыдущих. Подробнее читайте https://ru.wikipedia.org/wiki/Цепь_Маркова

$$K(U,H) = \frac{1}{V(U)^2 V(H)^2} \sum_{n}^{\infty} \mathbf{1}^T W^n \mathbf{1} \qquad (14)$$

Проблема такого подхода, однако, в том, что число маршрутов большой длины очень быстро возрастает и сумма расходится, стремясь к ∞. Чтобы этого избежать, необходимо понижать вклад в сумму от маршрутов большой длины. Это можно сделать с помощью введения в формулу (14) понижающего фактора, который удобно взять в виде $\lambda^n/n!$, где λ – какое-то число, обычно в диапазоне от 0 до 1, которое регулирует степень снижения вклада от длинных маршрутов. Тогда окончательное выражение для ядра будет следующим:

$$K(U,H) = \frac{1}{V(U)^2 V(H)^2} \sum_{n}^{\infty} \frac{\lambda^n}{n!} \mathbf{1}^T W^n \mathbf{1} \qquad (15)$$

Можно заметить, что полученная сумма в правой части уравнения очень похожа на разложение экспоненты в ряд Тейлора:

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots = \sum_{k}^{\infty} \frac{1}{k!} x^k \qquad (16)$$

Это позволяет упростить выражение (15) для ядра, избавившись от суммы:

$$K(U,H) = \frac{1}{V(U)^2 V(H)^2} \sum_{n}^{\infty} \frac{\lambda^n}{n!} \mathbf{1}^T W^n \mathbf{1} = \frac{1}{V(U)^2 V(H)^2} \mathbf{1}^T e^{\lambda W} \mathbf{1} \qquad (17)$$

Последняя формула является способом вычисления *экспоненциального графового ядра на случайных маршрутах*. Таким образом, вычисление числа общих маршрутов произвольной длины на двух графах можно провести без необходимости находить в явном виде все маршруты в графах. Вычисление графового ядра требует только проведения операций матричной алгебры.

Основное время, необходимое на вычисление графового ядра, уходит на нахождение матрицы, равной экспоненте матрицы смежности $e^{\lambda W}$, в связи с тем, что матрица *W* имеет довольно большой размер ($V(U)^2V(H)^2$ элементов). Тем не менее, используя свойства экспоненты, кронекеровского произведения и разреженность матрицы *W*, можно вычислять значения таких ядра достаточно эффективно. В то же время, при использовании легко вычисляемых фрагментных

дескрипторов расчет ядер сходства с использованием дескрипторного представления молекул все равно проходит значительно быстрее.

1.2.2.2. Отношения сходства для представления химических объектов в виде векторов дескрипторов

Отношения сходства для векторов бинарных дескрипторов

Представление химических объектов в виде векторов бинарных дескрипторов очень распространено в хемоинформатике. К этой категории относятся, в частности, «*молекулярные отпечатки пальцев*» (англ. fingerprints). В этом случае значение единицы в определенном бите (т.е. его «включенное» состояние) обозначает выполнение какого-либо условия, например, наличия определенной подструктуры в химическом объекте. Вектора бинарных дескрипторов называют также бинарными строками. Когда сравниваются вектора бинарных дескрипторов для химических объектов *A* и *B*, результат сравнения удобно представить в виде четырех чисел:

a - число бит, включенных для объекта A;

b - число бит, включенных для объекта B;

c - число бит, одновременно включенных как для объекта A, так и для объекта B;

d - число бит, одновременно выключенных как для объекта A, так и для объекта B.

Из всех индексов сходства между битовыми строками наибольшей популярностью (по крайней мере, в хемоинформатике) пользуются те из них, которые основаны на подсчете только чисел включенных битов, т.е. *a*, *b* и *c*. Одна из причин этого заключается в том, что включенный бит соответствует существенно более редкому, а потому и более важному явлению, например, наличию определенного фрагмента в химической структуре. Кроме того, более логично сравнивать химические объекты на основании того, что они имеют, чем на основе того, чего в них нет. В Табл. 1 представлены наиболее известные индексы сходства для векторов бинарных дескрипторов, основанные на подсчете включенных бит, а в Табл. 2 - расстояния между векторами бинарных дескрипторов в метрическом химическом пространстве.

Табл. 1. Индексы сходства для векторов бинарных дескрипторов, основанные на подсчете числа включенных бит

Название	Формула на языке		Ядро
	множеств	чисел битов	
коэффициент Жаккара (Jaccard) [23], коэффициент Танимото (Tanimoto)	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	$\frac{c}{a+b-c}$	+
коэффициент косинуса, коэффициент Карбо (Carbo)	$\frac{\lvert A \cap B \rvert}{\sqrt{\lvert A \rvert \cdot \lvert B \rvert}}$	$\frac{c}{\sqrt{a \cdot b}}$	+
коэффициент Дайса (Dice)	$\frac{2\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$	$\frac{2 \cdot c}{a+b}$	
коэффициент Кульчинского (Kulczinsky) [24]	$\frac{1}{2}\left(\frac{\lvert A \cap B \rvert}{\lvert A \rvert} + \frac{\lvert A \cap B \rvert}{\lvert B \rvert}\right)$	$\frac{c}{2}\left(\frac{1}{a} + \frac{1}{b}\right)$	
коэффициент Шимкевича-Симпсона (Szymkiewicz-Simpson) [25, 26]	$\frac{\lvert A \cap B \rvert}{min(\lvert A \rvert, \lvert B \rvert)}$	$\frac{c}{min(\lvert A \rvert, \lvert B \rvert)}$	
коэффициент Петке (Petke) [27]	$\frac{\lvert A \cap B \rvert}{max(\lvert A \rvert, \lvert B \rvert)}$	$\frac{c}{max(\lvert A \rvert, \lvert B \rvert)}$	
коэффициент Тверского (Tversky) [28]	$\frac{\lvert A \cap B \rvert}{\alpha\lvert A \backslash B \rvert + \beta\lvert B \backslash A \rvert + \lvert A \cap B \rvert}$	$\frac{c}{\alpha \cdot a + \beta \cdot b - c}$	-

Табл. 2. Расстояния для векторов бинарных дескрипторов, основанные на подсчете числа включенных бит

Название расстояния			Интервал	Метрика
	множеств	чисел битов		
Хэмминга (Hamming)	$\lvert A \cup B \rvert - \lvert A \cap B \rvert$	$a+b-2c$	[0,D]	+
Жаккара (Jaccard)	$1 - \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	$1 - \frac{c}{a+b-c}$	[0,1]	+

Если рассматриваются только включенные биты, то индексы сходства удобно также определять, пользуясь языком теории множеств. В этом случае при сравнении химических объектов A и B удобно рассматривать множество *A*, состоящее из битов, включенных для объекта A, и множество *B*, состоящее из битов, включенных для объекта B. Это дает возможность описывать индексы сходства и расстояния при помощи стандартных операций из теории множеств:

$\cap$ - пересечение множеств (множество объектов являющихся общими для двух заданных множеств);

$\cup$ - объединение множеств (множество объектов, присутствующих как минимум в одном из заданных множеств);

\ - разность множеств (множество объектов, присутствующих в первом множестве, но отсутствующих во втором);

$|\cdot|$ - мощность множества (т.е. число элементов в нем).

В Табл. 1 и Табл. 2 представлены формулы для индексов сходства и расстояний, записанные на языке теории множеств. Некоторые из индексов сходства для бинарных дескрипторов являются ядрами и поэтому могут быть использованы в ядерных методах машинного обучения, таких как метод опорных векторов, Гауссовы процессы и ядерный метод гребневой регрессии. Наиболее популярным в хемоинформатике из таких индексов сходства является, вне всякого сомнения, коэффициент Танимото, который используется в абсолютном большинстве исследований в сочетании с молекулярными отпечатками пальцев (фингерпринтами) для оценки сходства химических объектов и для организации поиска по подобию в химических базах данных. Большинство из используемых расстояний между битовыми строками, такие как расстояние Хэмминга и Жаккара, являются метриками, т.е. для них справедливо неравенство треугольника. Заметим, что индекс Танимото и расстояние Жаккара тесно связаны между собой: их сумма всегда равна единице. Большинство типов индексов сходства и расстояний являются симметричными, т.е. их значение не меняется при перестановке химических объектов. Наиболее известным исключением из этого правила является индекс Тверского (Tversky), значение которого разное в зависимости от того, какой объект мы считаем первым, а какой – вторым. В последнее время индекс Тверского нашел широкое применение в хемоинформатике при организации поиска по сходству вследствие естественной асимметрии этой процедуры, поскольку роль молекулы-запроса отличается от роли молекулы из базы данных, по которой ведется поиск.

Отношения сходства между векторами произвольных дескрипторов

В Табл. 3 и Табл. 4 представлены основные меры сходства и расстояния для векторов произвольных дескрипторов. Под произвольными дескрипторами понимаются в данном случае дескрипторы, принимающие вещественные[1] и целые значения (как правило, натуральные, т.е. целые неотрицательные), тогда как бинарные дескрипторы обычно выделяют в рассмотренную выше отдельную категорию, поскольку в некоторых важных случаях их меры сходства обладают разными свойствами. Например, индекс Танимото, который является корректным ядром для векторов бинарных дескрипторов, уже не является корректным ядром для векторов вещественных и целочисленных дескрипторов, и поэтому его неправильно использовать в ядерных методах машинного обучения. В отличие от бинарных векторов, для которых наиболее популярной (по крайней мере в хемоинформатике) мерой сходства является коэффициент Танимото, для произвольных векторов эту роль обычно играет евклидово расстояние, а в отдельных работах также используется расстояние по Махаланобису, основанное на использовании произвольной метрики. Целочисленные значения обычно принимают дескрипторы, представляющие собой частоты либо индикаторы встречаемости каких-либо фрагментов в химической структуре. Происхождение вещественных значений весьма разнообразно – такие дескрипторы могут отвечать за величину какого-то измеряемого (например, коэффициент распределения вода-октанол) или вычисляемого (например, индекс Рандича) свойства химического объекта. Вектора дескрипторов, представленных натуральными и вещественными числами, используются в расчете индексов сходства и расстояний с использованием одних и тех же формул. Формулы для расчета мер сходства для двух объектов A и B, представленных векторами $\boldsymbol{a} = \{a_1, a_2, \dots a_D\}$ и $\boldsymbol{b} = \{b_1, b_2, \dots b_D\}$ приведены в Табл. 3, а для расчета расстояний между этими объектами – в Табл. 4. Также указано, какие из мер сходства являются ядрами и какие из расстояний – метриками.

[1] Для простоты мы не обсуждаем различий между вещественными и рациональными числами, поскольку это не влияет на способы использования дескрипторов.

Табл. 3. Меры сходства для векторов произвольных дескрипторов

Название индекса сходства	Формула индекса сходства	Интервал	Ядро
коэффициент косинуса (Cosine), коэффициент корреляции, коэффициент Карбо (Carbo)	$\frac{< \mathbf{a},\mathbf{b} >}{\sqrt{< \mathbf{a},\mathbf{a} > \cdot < \mathbf{b},\mathbf{b} >}}$	[-1,1]	+
коэффициент Ходжкина (Hodgkin) [29]	$\frac{2 < \mathbf{a},\mathbf{b} >}{< \mathbf{a},\mathbf{a} > + < \mathbf{b},\mathbf{b} >}$	[-1,1]	
коэффициент Петке (Petke) [27]	$\frac{< \mathbf{a},\mathbf{b} >}{max(< \mathbf{a},\mathbf{a} >, < \mathbf{b},\mathbf{b} >)}$	[-1,1]	
Скалярное произведение	$< \mathbf{a},\mathbf{b} >$	[-∞,∞]	+
Полиномиальное ядро	$(1+< \mathbf{a},\mathbf{b} >)^p$	[-∞,∞] или [0,∞]	+
Гауссово ядро	$\exp(-\gamma < \mathbf{a}-\mathbf{b},\mathbf{a}-\mathbf{b} >)$	[0,1]	+
Индекс Танимото для произвольных векторов	$\frac{< \mathbf{a},\mathbf{b} >}{< \mathbf{a},\mathbf{a} > + < \mathbf{b},\mathbf{b} > - < \mathbf{a},\mathbf{b} >}$	$[-\frac{1}{3}, 1]$	-

Табл. 4. Расстояния для произвольных векторов

Название расстояния	Формула для расстояния	Метрика
Эвклидово (Euclidean)	$\sqrt{\sum_{i=1}^{D}(a_i - b_i)^2}$	+
По Манхэттену (Manhattan)	$\sqrt{\sum_{i=1}^{D}\lvert a_i - b_i\rvert}$	+
Минковского (Minkowski)	$\left(\sum_{i=1}^{D}\lvert a_i - b_i\rvert^p\right)^{\frac{1}{p}}$	+ (p≥1) - (p<1)
Махаланобиса (Mahalanobis)	$\sqrt{\sum_{i=1}^{D}\sum_{j=1}^{D}(a_i - b_i)m_{ij}(a_j - b_j)}$ m_{ij} – элемент матрицы метрики	+

1.2.2.3. Отношения сходства для представления химических объектов в виде функций

С точки зрения физики молекулы представляют собой статистические распределения: электроны и даже ядра атомов распределены в пространстве в соответствии с некоторой функцией (электронной плотностью в случае электронов). Такая функция характеризует распределение некоторого поля в пространстве вокруг молекулы и называется молекулярным полем (англ. *molecular field*). Она может быть задана в аналитическом виде: например, электронная плотность в методах квантовой химии может быть аппроксимирована как комбинация гауссовых базисных функций. Кроме того, функция может быть задана численно, то есть указаны ее значения для набора точек в узлах воображаемой решетки, заданной в пространстве вокруг молекулы. Именно этот способ представления функций, описывающих молекулярные поля, которые используют большинство методов 3D QSAR (см. раздел 4.1 пособия 3). Функции, которые могут быть представлены таким образом, довольно разнообразны: распределения электронной плотности, электростатического или ван-дер-ваальсова потенциала вокруг молекулы, энергии взаимодействия с пробными атомами или молекулами, помещенными в данную точку пространства. Меры сходства в этом случае могут быть вычислены аналитически путем интегрирования произведений таких функций по физическому пространству. Наиболее часто используемые из них представлены в Табл. 5. В приведенных формулах функция $f_a(\mathbf{r})$ относится к молекуле a, а $f_b(\mathbf{r})$ – к молекуле b.

В разделе 5.5.1 настоящего пособия будет показано, как распределение электронной плотности может быть использовано для оценки сходства молекул. В разделе 5.5.2.4 будет объяснено, как можно сравнить форму молекул с использованием функций Гаусса, а в разделе 5.5.3.1 – как вычисляется сходство молекул с использованием распределения электростатического и ван-дер-ваальсова потенциала.

Табл. 5. Меры сходства для выраженных в виде функций молекулярных полей

Название индекса сходства	Формула индекса сходства	Интервал	Ядро
коэффициент Карбо (Carbo)	$\frac{\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}{\sqrt{\iiint f_a(\mathbf{r})f_a(\mathbf{r})d\mathbf{r}}\cdot\sqrt{\iiint f_b(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}}$	[-1,1]	+
коэффициент Ходжкина (Hodgkin) [29]	$\frac{2\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}{\iiint f_a(\mathbf{r})f_a(\mathbf{r})d\mathbf{r}+\iiint f_b(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}$	[-1,1]	
Коэффициент Петке (Petke) [27]	$\frac{\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}{\max\left(\iiint f_a(\mathbf{r})f_a(\mathbf{r})d\mathbf{r},\iiint f_b(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}\right)}$	[-1,1]	
Индекс Танимото (Tanimoto) для полей	$\frac{\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}{\iiint f_a(\mathbf{r})f_a(\mathbf{r})d\mathbf{r}+\iiint f_b(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}-\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}$	[0,1] при $f(\mathbf{r})>0$	
Индекс Тверского (Tversky) для полей	$\frac{\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}{\alpha\iiint f_a(\mathbf{r})f_a(\mathbf{r})d\mathbf{r}+\boldsymbol{\beta}\iiint f_b(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}-\iiint f_a(\mathbf{r})f_b(\mathbf{r})d\mathbf{r}}$	[0,1] при $f(\mathbf{r})>0$ и $\alpha+\beta=1$	

2. ОПИСАТЕЛЬНЫЙ АНАЛИЗ ХИМИЧЕСКОГО ПРОСТРАНСТВА

Химическое пространство считается заданным, когда определены отношения сходства между химическими объектами (химическими соединениями, смесями, реакциями, материалами и др.). Применение концепции химического пространства в хемоинформатике связано с анализом распределения в нем объектов и поиском закономерностей. Как было отмечено в предыдущей главе, объекты химического пространства могут быть представлены графами, строками чисел и функциями. Для каждого типа объектов разработаны собственные инструменты анализа их распределения в химическом пространстве.

2.1. ХИМИЧЕСКОЕ ПРОСТРАНСТВО ГРАФОВ

В принципе, любое множество молекулярных графов формирует дискретное топологическое пространство. Его топология формально определяется множеством всех возможных подмножеств, а простейшая метрика определяет расстояние между графами как равное 0, если химические объекты эквивалентны (в частности, если соответствующие молекулярные графы изоморфны), и 1 - если неэквивалентны. Другие, более практически полезные способы вычисления отношений сходства между графами обсуждались ранее в разделе 1.2.2.1. Некоторые из них, основанные, например, на представлении графов с использованием ядер молекулярного сходства, позволяют использовать для визуализации некоторые методы машинного обучения (например, ядерный вариант метода главных компонент, KernelPCA [30]). Кроме того, вычисление расстояний между объектами, представленными в виде графов, позволяет использовать для них методы визуализации, основанные на графах соседства, который будет рассмотрен в разделе 2.4, а также методы визуализации, основанные на вычислении расстояний между объектами, например, *картирования по Сэммону* (англ. Sammon mapping) [31].

Далее рассмотрим способы анализа химического пространства графов, которые специфичны для данного типа химических объектов. Существует три типа подходов к представлению набора молекулярных графов и навигации в пространстве графов: (*а*) подструктурный; (*б*) надструктурный; и (*в*) мутационный. Кроме того, с работой в

химическом пространстве графов непосредственно связан описанный ниже метод молекулярных пар соответствия, широко используемый в настоящее время при проведении анализа «структура-активность» для дизайна новых химических соединений.

2.1.1. Подструктурный подход. Молекулярные каркасы и остовы

В *подструктурном* подходе строится специальный «навигационный» граф, который может быть использован как для визуализации баз данных, так и для поиска неисследованных областей химического пространства. В навигационном графе вершины соответствуют молекулярным графам, а ребра - переходам между ними в соответствии с определенными правилами. Бемис и Мурко (Bemis, Murcko) рассмотрели переходы от помеченного графа (полной химической структуры) к непомеченному, а от него – к графам без боковых цепей [32, 33]. Они ввели концепцию молекулярных *каркасов* (англ. frameworks) [32, 33], которые позволяют организовать структурные данные путем отнесения атомов к кольцам, линкерам и боковым цепям (Рис. 7). Впоследствии их стали также называть остовами (скаффолдами). С помощью этого приема даже базы данных, содержащие огромное число химических соединений, могут быть описаны при помощи сравнительно небольшого числа каркасов. В частности, в своих работах [32, 33] Бемис и Мурко проводили анализ частоты встречаемости каркасов в базе DrugBank и показали, что большинство существующих лекарств покрываются небольшим числом каркасов. Это значит, что структурное разнообразие лекарств относительно невелико.

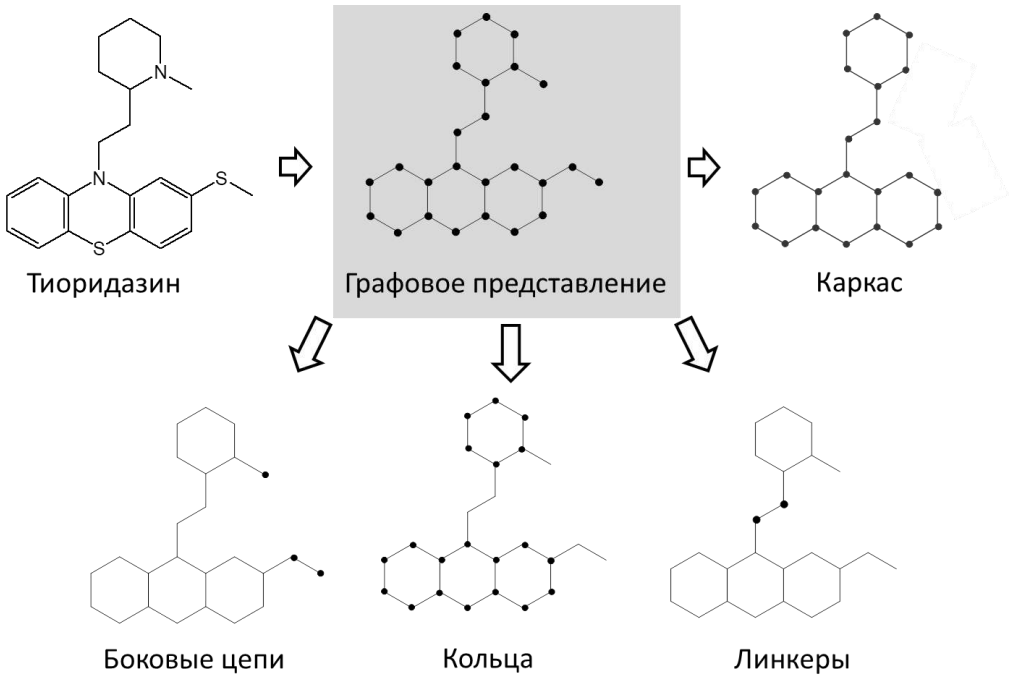


Рис. 7. Подход Бемиса-Мурко к выделению каркасов молекул. Для этого структура молекулы (приведена молекула лекарства тиоридазина) кодируется в граф с потерей информации о типах атомов и связей. Каркасом считается подструктура, которая содержит все кольца и соединяющие их цепочки атомов - линкеры. В каркас не входят боковые цепи.

В «дереве остовов» («scaffold tree»), предложенном Шуфенхауэром (Schuffenhauer) с соавт. [34, 35], переходы разрешены между молекулярным графом и его подграфом. В отличие от подхода Бемиса и Мурко, в нем рассматриваются помеченные молекулярные графы, а атомы кислорода карбонильных групп, атомы углерода которых входят в кольцо, считаются принадлежащими этому кольцу. Структуры, полученные удалением боковых цепей, называются *остовами* (*скаффолдами*, англ. *scaffolds*). Авторы разработали набор правил, которые разбивают остовы на структуры меньшего размера (Рис. 8). Правила соответствуют синтетической сложности введения того или иного фрагмента в остов, причем родительский остов итеративно упрощается удалением «младшего» кольца.

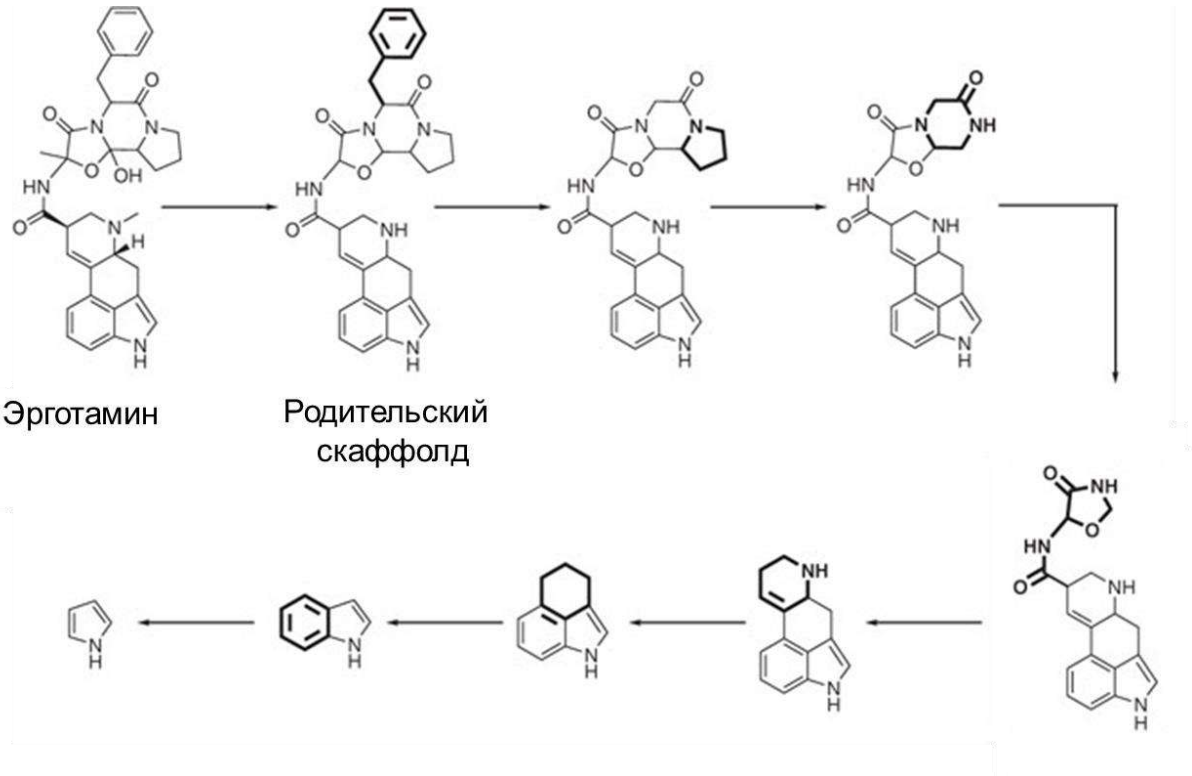


Рис. 8. Пример создания иерархии остовов (скаффолдов) для молекулы лекарственного препарата эрготамина. Удаляемая (младшая) группировка в остове выделена жирными линиями. Рисунок адаптирован из статьи [34] и приводится с разрешения издательства. Copyright (2007) American Chemical Society.

Ниже приведены предложенные в упомянутых публикациях правила получения следующего по иерархии остова, причем правило, стоящее ниже в приведенном списке, применяется только тогда, когда вышележащее не позволяет выделить младшую группу:

1. Удаляются 3-членные гетероциклы
2. Не удаляются кольца, содержащие более 12 атомов
3. Удаляются циклы, соединенные более длинными линкерами
4. Не удаляются три- и полициклы с мостиковыми атомами, спироциклы и нелинейные конденсированные циклы. Для этого вычисляется характеристика Δ = (число связей, принадлежащий более, чем 1 циклу) – (число циклов) + 1. Младшим считается тот фрагмент, удалением которого получается остов с максимальным $|\Delta|$ (модулем от Δ).
5. Системы с мостиковыми атомами считаются старше спироциклов. Другими словами, если получающиеся остова имеют одинаковые $|\Delta|$, то удаляется кольцо, которое приводит к получению остова с отрицательным значением Δ.

6. Удаляются кольца размером 3 > 5 > 6
7. Ароматическая система не разрушается, если получившаяся структура не ароматична.
8. Удаляются кольца с меньшим число гетероатомов.
9. Если число гетероатомов в циклах одинаково, считаются более старшими гетероциклы с S > N > O
10. Сначала удаляются циклы меньшего размера
11. В системах, содержащих ароматические и неароматические фрагменты, сначала удаляются ароматические.
12. Удаляются циклы, в который линкер связан с гетероатомом.
13. Остаются остовы, чей канонический SMILES является младшим в алфавитном порядке. Это правило служит для разрешения оставшихся конфликтов.

Было показано, что этот тип навигационных графов позволяет выявлять части химического пространства, богатые активными структурами. благодаря чему он может служить ценным инструментом для дизайна биологически активных соединений [36]. Идея «дерева остовов» легла в основу популярной свободно-распространяемой программы «Scaffold Hunter» [37], позволяющей в интерактивном режиме осуществлять навигацию по химическому пространству графов, выявлять скрытые закономерности «структура-свойство» и находить области, насыщенные активными соединениями, Рис. 9.

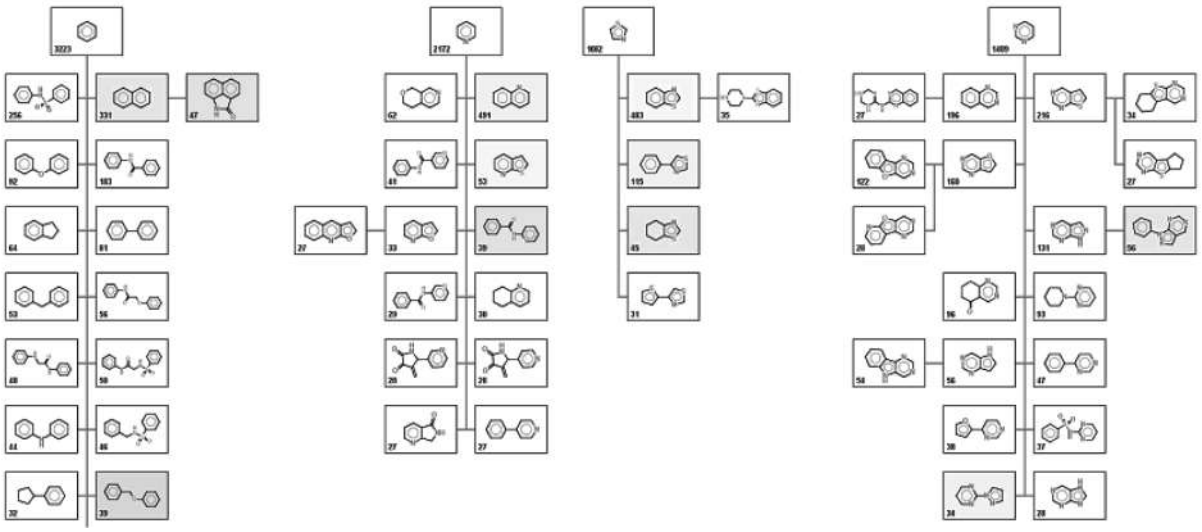


Рис. 9. Пример дерева остовов для некоторых ингибиторов пируват-киназы, генерируемого программой Scaffold Hunter. Более темным цветом выделены остовы, соответствующие более активным соединениям. Рисунок адаптирован из статьи [34]. Copyright (2007) American Chemical Society.

Для визуализации химического пространства, образованного более узкими наборами соединений с одинаковым остовом, но с разными заместителями, были разработаны специальные «комбинаторные графы аналогов» [38], позволяющие иерархически упорядочивать соединения в соответствии с комбинациями заместителей. Будучи аннотированы с помощью индексов связи структура-активность SARI (см. раздел 2.2.3) [39], они позволяют анализировать гладкость зависимости «структура-активность» на уровне функциональных групп. Эти графы позволяют идентифицировать малоисследованные области химического пространства, а также выявлять комбинации заместителей, определяющие биологическую активность. Альтернативным способом визуализации структур с одинаковым остовом являются «карты SAR», предложенные Аграфиотисом с соавт. [40]. В них каждый набор соединений с одинаковым остовом представлен в виде матрицы, каждая ячейка которой соответствует уникальной комбинации заместителей. Ячейки «карты SAR» окрашены в соответствии с уровнем изучаемой активности химического соединения, что позволяет быстро идентифицировать заместители, в наибольшей степени влияющие на биологическую активность.

Поллок с соавт. [41] для идентификации топологии остовов ввели специальный граф, позволяющий при минимальном числе узлов и ребер полностью описать систему циклов в молекуле. Введение этого графа позволило систематически сгенерировать и проанализировать все топологии остовов с числом циклов до восьми, что обеспечивает покрытие значительной доли химического пространства для малых молекул [41]. Это позволило провести анализ распределений топологий остовов на нескольких широко известных базах данных с очень большим числом соединений, как реальных, так и виртуальных, и выявить множество закономерностей [42]. Подобные графы, идентифицирующие топологию остовов, составляют первую ступень в иерархической классификации молекул в химических базах [42].

2.1.2. Надструктурный подход

В *надструктурном* подходе каждый индивидуальный молекулярный граф рассматривается как подграф общего суперграфа. [43] Хотя этот подход ограничен сравнительно узкими наборами сходных по строению химических соединений, на его основе разработаны эффективные способы построения моделей QSAR:

позиционный анализ Мажи (Magee) [44, 45], система DARC/CALPHI [46], MTD-PLS [47-49], метод MFTA Палюлина с соавт. [43, 50].

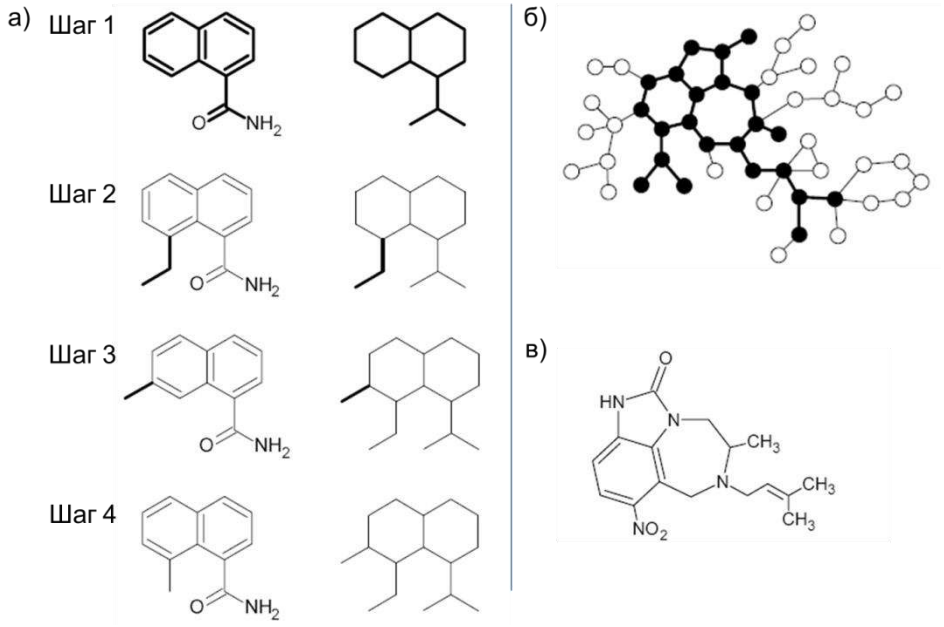


Рис. 10. (а) Построение молекулярного суперграфа в методе MFTA для 4 структур. Слева приведены предъявляемые структуры, справа – получаемый на текущей итерации суперграф. Жирными линиями выделены ребра, соединяющие добавляемые на текущей итерации вершины графа. (б) Пример суперграфа, построенного для ряда бензодиазепиновых производных. На суперграфе выделена подструктура, соответствующая молекуле, приведенной на рисунке (в).

Рассмотрим подход MFTA (англ. *Molecular Field Topology Analysis*), развитый В.А. Палюлиным и Е.В. Радченко под руководством Н.С. Зефирова [43, 50]. При создании суперграфа молекулы базы данных последовательно накладываются друг на друга, причем на каждом этапе этой процедуры определяется максимальная общая подструктура между текущим суперграфом и текущей молекулой. Метки атомов (тип химического элемента) и типы связей при этом не учитываются, и поэтому получающийся суперграф представляет собой простой непомеченный граф. Если в текущей молекуле содержится фрагмент, отсутствующий в суперграфе (Рис. 10а), то в суперграф дополняется отсутствующими вершинами. Часто при этом могут возникать ситуации, когда существует несколько вариантов вложения молекулы в суперграф и, следовательно, имеется неоднозначность в том, как можно обновить суперграф. Тогда перебираются все варианты, и оставляется тот, в котором локальные свойства атомов и их окружений (заряды, ван-дер-ваальсовы радиусы

и пр.) в максимальной общей подструктуре молекулы и суперграфа предельно близки. Формула, которая ранжирует возможные вложения, описана в работе [50]. В ней глубина окружения и учитываемые характеристики атомов являются настраиваемыми параметрами метода. Пример получающегося при этом суперграфа представлен на Рис. 10б.

Применение надструктурного подхода в QSAR-моделировании обусловлено тем, что для каждой химической структуры занятость вершин суперграфа, либо физико-химические характеристики атомов, соответствующих вершинам суперграфа, образуют вектор дескрипторов фиксированного размера, который можно использовать при построении моделей QSAR в качестве входа для методов машинного обучения. Для каждой молекулы, свойство которой нужно предсказать (например, приведенной на Рис. 10в), находится ее вложение в суперграф (выделено на Рис. 10б). Каждой вершине суперграфа соответствует вектор дескрипторов, характеризующих какие-либо свойства атома, вкладывающегося в данную вершину. Если атом вкладывается в данную вершину (черные на Рис. 10б), то берутся характеристики данного атома, если нет (белые на Рис. 10б), то считаются, что вектор заполнен нулями. Конкатенацией векторов дескрипторов, соответствующих отдельным вершинам суперграфа, получается вектор дескрипторов для молекулы, который может использоваться в QSAR-моделировании. Некоторые примеры применения данного подхода приведены в работе [50].

2.1.3. Подход, основанный на мутациях

Альтернативный подход к навигации в химическом пространстве графов, основанный на использовании *мутаций,* был предложен Ван Дюрсеном (van Deursen) с соавт. [51]. В его рамках химическое пространство представляется как граф, вершины которого соответствуют индивидуальным молекулам, а ребра – возможным модификациям химической структуры (мутациям), таким как: (а) изменение типа атома; (б) инверсия стереохимической конфигурации при хиральных центрах; (в) удаление либо добавление атома; (г) образование, разрыв или изменение порядка связи; (д) добавление либо удаление ароматического кольца. Передвигаясь в представленном таким образом пространстве от одной активной структуры к другой, можно обнаружить и собрать вдоль пути определенное число новых структур, от которых можно также ожидать проявления той же самой биологической активности. В подходе, предложенном Бишопом

(Bishop) с соавт. [52], в качестве структурных мутаций выступает набор известных органических реакций. Этот подход позволил авторам предложить набор «самых полезных соединений», из которых большинство остальных может быть синтезировано.

2.1.4. Анализ пар соответствия молекул

2.1.4.1. Основы анализа пар соответствия молекул

Исследователи всегда больше склонны доверять таким подходам к молекулярному дизайну, которые напоминают им методы исследования, которыми они сами привыкли пользоваться, поскольку в их основе лежит опыт, накопленный целыми поколениями исследователей в течение многих лет. В частности, в процессе дизайна новых молекул исследователи часто пользуются множеством эмпирических закономерностей (назовем их правилами) о том, что введение в молекулу определенной группы атомов приводит к определенным изменениям в физико-химических свойствах (например, в растворимости или липофильности) и в биологической активности разрабатываемых химических соединений. Такие правила дают возможность сразу определить, какие изменения надо внести в молекулу, чтобы сделать разрабатываемое химическое соединение более активным и/или оптимизировать его свойства. Такой подход к молекулярному дизайну был формализован и лег в основу концепции анализа *пар соответствия молекул* (англ. Matched Molecular Pairs - *MMP*) [53-58].

Термин «*пары соответствия молекул*» был впервые введен в 2005 г. Кенни (Kenny) и Садовски (Sadowski) [53] и обозначает **пару молекул, различающихся осуществленной в одном месте четко определенной трансформацией**, которую можно связать с разницей в значениях их свойств [57]. Такое определение является, однако, очень общим и поэтому требует конкретизации. Обычно полагается, что две молекулы различаются лишь одной связной подструктурой с небольшим числом атомов, тогда как общая их часть содержит основную долю атомов, причем эти две части молекул соединены друг с другом в одной (чаще всего), двух (реже) или трех (еще реже) точках. Если соединение происходит в одной точке, то такая трансформация носит характер замены функциональной группы. На Рис. 11 приведена в качестве примера пара соответствия молекул, различающихся заменой атома хлора в качестве заместителе в ароматическом кольце на карбоксильную группу, чему соответствует приведенная справа

трансформация, показанная как с помощью структурной диаграммы, так и с помощью строки SMIRKS. На Рис. 12 приведен пример пары соответствия молекул, отличающихся заменой остатка фурана на бензольный цикл, с двумя точками присоединения.

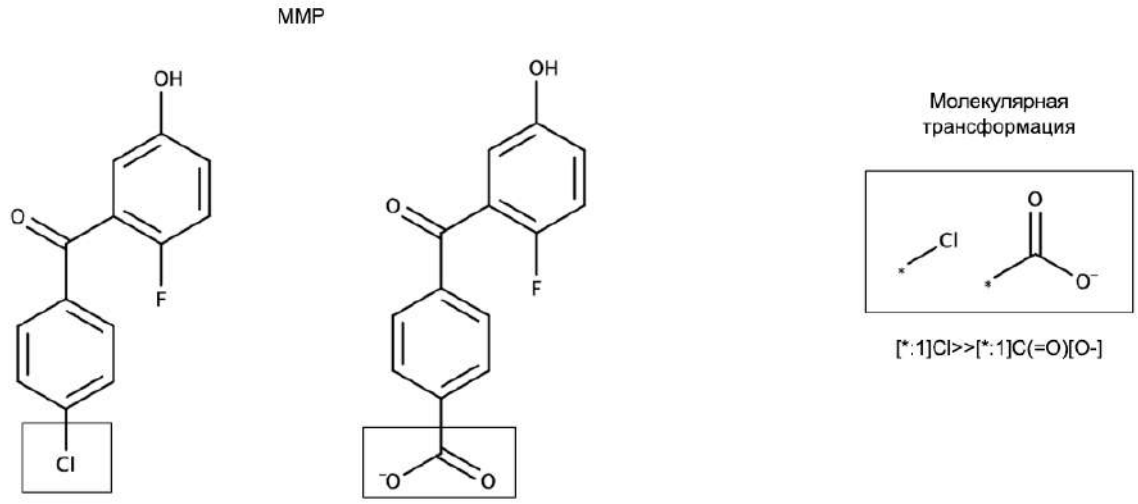
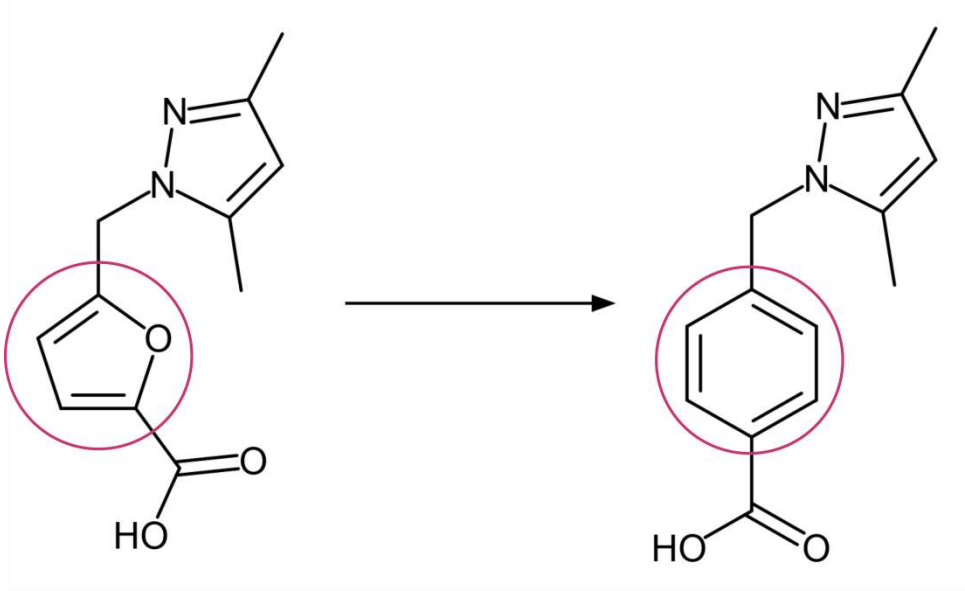


Рис. 11. Пара соответствия молекул с одной точкой присоединения, отличающихся заменой атома хлора (–Cl) на карбоксильную группу ($–CO_2^-$) (слева и в центре, соотвественно), а такжесоответствующая ей трансформация (справа), обозначенная при помощи структурной диаграммы и строки SMIRKS.



SMIRKS: [*:1]c1ccc([*:2])o1>>[*:1]c1ccc([*:2])cc1

Рис. 12. Пара соответствия молекул с двумя точками присоединения, отличающаяся заменой остатка фурана на бензольный цикл.

С каждой парой соответствия связана величина изменения значения свойства, вычисляемая как разница между значениями этого свойства для пары молекул: ΔpAct = pAct2 – pAct1. Если оказывается, что одна и та же трансформация всегда или в большинстве случаев приводит к аналогичному эффекту на данное свойство, то такая трансформация считается *существенной* (англ. significant transformation), и на ее основе может быть сформулировано правило: такая-то трансформация приводит к такому-то эффекту. Набор таких правил, полученный в результате анализа базы данных, образует набор правил, который можно использовать для дизайна новых молекул с целью оптимизации набора свойств. Например, для того чтобы узнать, как изменить структуру молекулы, чтобы сделать ее более активной, надо найти в сформированном таким образом наборе такое правило, которое применимо к данной молекуле и которое приводит к необходимому изменению активности. Применение соответствующей трансформации к молекуле сразу же приводит к новой молекуле, которую можно синтезировать с целью повышения активности. Если же надо, например, повысить растворимость в воде без изменения активности, то надо найти в сформированном наборе такое правило, которое было бы применимо к данной молекуле, предсказывало бы повышение растворимости в воде, но при этом, однако, должно отсутствовать правило, указывающее на то, что соответствующая трансформация влияет на активность молекулы. Таким образом, набор правил, сформулированный при анализе пар соответствия молекул (пар ММР), является ценным инструментом для дизайна биологически активных молекул.

Набор правил удобно организовать в виде графа трансформаций, вершины которого соответствуют модифицируемым подструктурам (обычно заместителям), а ребра – существенным трансформациям между ними, причем ребра направлены в сторону повышения активности. В качестве примера, на представлен полученный в работе [59] граф трансформаций для водной токсичности (англ. aquatic toxicity).

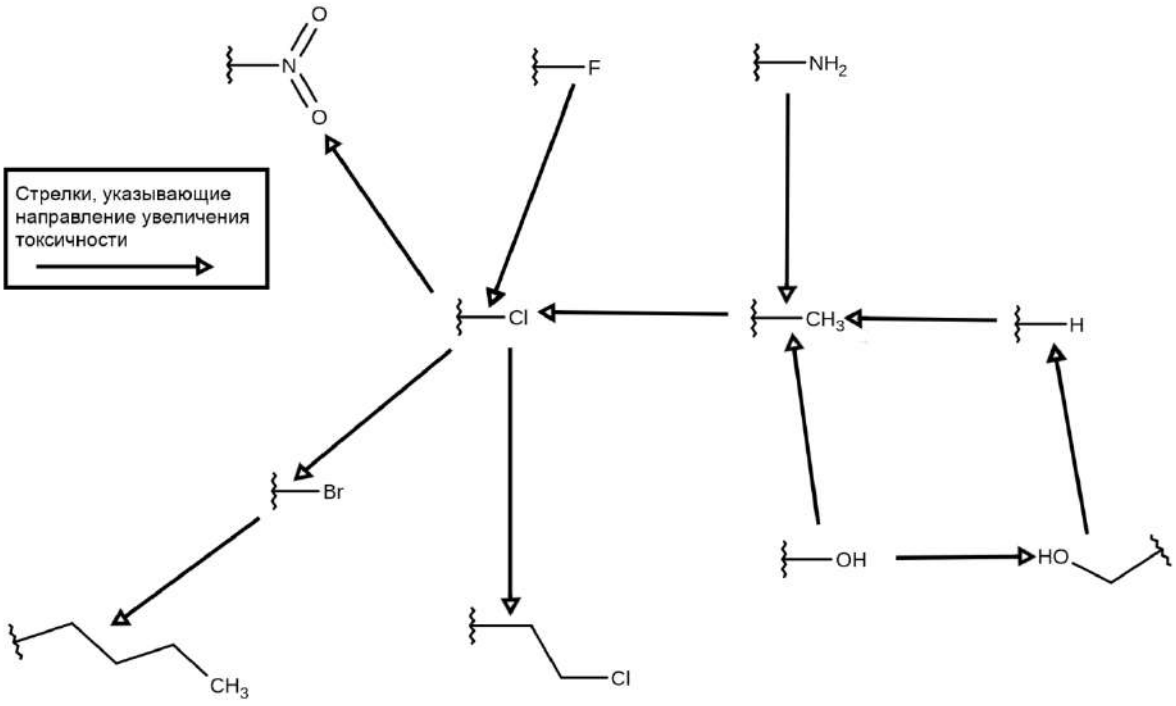


Рис. 13. Пример графа трансформаций для водной токсичности. Рисунок из публикации [59] с открытым доступом.

Благодаря представлению набора правил в виде графа трансформаций удается выявить дополнительные закономерности, полезные для оптимизации структур молекул. Например, как видно на [59], наличие брома в качестве заместителя приводит к высокой токсичности, потому что он находится ближе к концу путей в направленном графе, тогда как наличие в этом же положении гидроксильной группе приводит к низкой токсичности, поскольку он находится в начале путей в графе трансформаций.

2.1.4.2. Основные статистические характеристики в анализе пар соответствия молекул

Для того, чтобы определить, является ли трансформация, описывающая пару соответствия молекул, существенной и, следовательно, может быть использована для формулировки правила, необходимо провести статистический анализ. С этой целью для данной трансформации надо найти в базе данных все пары MMP, которые она описывает. В том случае, если выбранное свойство описывается количественной мерой pAct, то набор величин его изменения при данной трансформации может быть представлен в виде гистограммы. В качестве примера на Рис. 14 приведена взятая из статьи [54] гистограмма изменения величины pIC_{50} для EDFR (рецептор

эпидермального фактора роста), происходящего при введении гидроксильной группы в ароматическое кольцо.

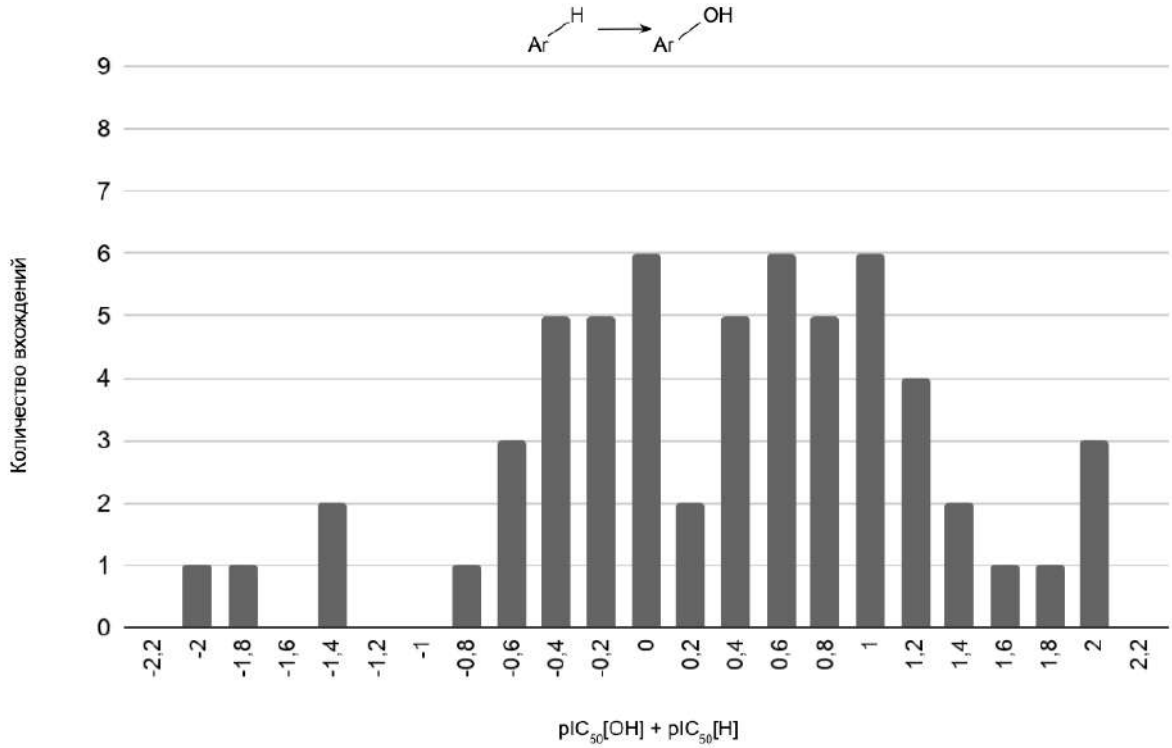


Рис. 14. (вверху) Рассматриваемая трансформация – введение гидроксильной группы в ароматическое кольцо. (внизу) Гистограмма изменения значения свойства pIC_{50} для всех пар MMP, соответствующих этой трансформации. Рисунок адаптирован из публикации [54] и приводится с разрешения издательства. Copyright (2011) American Chemical Society.

Рассматриваемая трансформация (замена атома водорода на гидроксильную группу в ароматическом кольце) обнаружена в 59 парах MMP в подвергнутой анализу базе данных. Легко видеть, что данная трансформация может приводить как к повышению (положительные значения разности), так и к понижению (отрицательные значения разности) активности молекулы. Тем не менее, заметно, что в большинстве случаем имеет место повышение активности. Для того, чтобы понять, является ли трансформация существенной и поэтому может ли быть подобная закономерность использована включена в набор правил, предназначенных для дизайна новых молекул, необходимо определить, является ли она статистически значимой. С этой целью может быть использован аппарат проверки статистических гипотез. Это может быть сделано несколькими способами. В том случае, когда величина изменения свойства описана в базе данных количественной величиной (в данном

случае ΔAct_i для i-ой пары ММР), можно вычислить значения среднего изменения значения свойства на N парах ММР (*Mean*):

$$Mean = \frac{1}{N}\sum_{i=1}^{N}\Delta Act_i \quad (18)$$

стандартного отклонения для ΔAct_i (*SD*):

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\Delta Act_i - Mean)^2} \quad (19)$$

и стандартной ошибки средней величины (*SEM*):

$$SEM = \frac{SD}{\sqrt{N}} \quad (20)$$

В этом случае статистическая значимость правила «данная трансформация приводит к росту активности» может быть оценена при помощи одновыборочного *t*-критерия Стьюдента (англ. one-sample t-test) по формуле:

$$t = \frac{Mean}{SEM} > t_{crit}(N-1, 0.05) \approx 2.0 \quad (21)$$

Поскольку в большинстве практически важных случаев критическое значение t_{crit} при стандартном уровне значимости 0.05 примерно равно 2.0, то можно считать, что такое правило будет статистически значимо, если средняя величина изменения свойства (Mean) больше, чем в два раза выше, чем значение стандартной ошибки для него (*SEM*). При отрицательных значениях *t*-критерия, по модулю превышающих 2.0, статистически значимым становится правило «данная трансформация приводит к понижению активности».

Если принимать во внимание только знак изменения значения свойства (например, активность возрастает либо уменьшается), то для проверки статистической значимости того, что трансформация является существенной и может быть сформулировано правило «данная трансформация чаще приводит к повышению, чем к понижению значения данного свойства (активности)», можно рассчитать *p*-величину для биноминального распределения и сравнить со стандартным уровнем значимости 0.05:

$$p - value = F_{bin}(\min(N_+, N_-)\,; N; 0.5) < 0.05 \qquad (22)$$

где N_+ - число соответствующих данной трансформации пар MMP, для которых наблюдается увеличение величины свойства; N_- - число соответствующих данной трансформации пар MMP, для которых наблюдается уменьшение величины свойства; N – общее число пар MMP, соответствующих данной трансформации, а кумулятивная функция распределения для биноминального распределения F_{bin} вычисляется по формуле:

$$F_{bin}(k; N; p) = \sum_{i=0}^{\lfloor k \rfloor} \binom{N}{i} p^i (1-p)^{N-i} \qquad (23)$$

Если p-значение окажется ниже 0.05, то соответствующую трансформацию можно считать существенной и соответствующее правило может быть включено в набор для последующего использования для дизайна новых молекул.

2.1.4.3. Иерархия структурных контекстов в анализе пар соответствия молекул

При анализе пар соответствия молекул важно учитывать структурный контекст трансформации, то есть окружение замещаемых групп. В качестве примера на Рис. 15 приведена рассмотренная в статье [54] одна и та же трансформация замены атома водорода на гидроксильную группу в разных структурных контекстах, образующих иерархию. В качестве анализируемого свойства рассматривается величина pIC_{50} лиганда по отношению к EDFR (рецептор эпидермального фактора роста). Для каждого из структурных контекстов приведены гистограмма распределения величин pIC_{50}, число пар MMP (N) и значение рассмотренных выше статистических характеристик *Mean* и *SD* (обозначена как σ). Верхушку иерархии образует максимально обобщенный контекст, когда рассматриваются все трансформации, сводящиеся к замене атома водорода на гидроксильную группу. Такой трансформации соответствует наибольшее число пар соответствия в данной иерархии – 75 примеров. Для него значение t-критерия равно 2.4. Поскольку оно превышает пороговое значение 2.0, то закономерность «замена атома водорода на гидроксил приводит к повышению активности» является статистически значимой на уровне 0.05 и поэтому, в принципе, может быть включена в набор правил. Второй уровень иерархии образуют

трансформация «замена атома водорода у ароматического атома на гидроксильную группу», соответствующая 59 парам ММР, (слева) и «замена атома водорода у алифатического атома на гидроксильную группу», соответствующая 16 парам ММР (справа). В данном случае структурными контекстами являются «у ароматического атома» и «у алифатического атома». Для первого из них значение *t*-критерия равно 3.18, что превышает 2.4 и, тем более, 2.0, и поэтому правило «замещение на гидроксильную группу в ароматическом кольце приводит к повышению активности» может быть включено в набор для последующего молекулярного дизайна с большей надежностью, чем для более широкого контекста. Соответствующее правило для этого более специфического контекста является более надежным, однако и покрытие (т.е. число пар ММР, к которым оно применимо) нескольку ниже. Для второго из них, описывающего замещение при алифатическом атоме углерода, значение *t*-критерия -1.14 и поэтому для него не может быть сформулировано статистически значимое правило. Третий уровень иерархии образуют более специфические структурные контексты. Например, трансформация слева внизу на Рис. 15 описывает замену атома водорода у атома углерода замещенного анилина в пара-положении к группе $–NH_2$ на гидроксильную группу. Для него, однако, есть только 5 примеров в базе данных, и значение *t*-критерия всего лишь -0.77, и поэтому статистически значимого правила не может быть сформулировано. То же относится и к другой приведенной на рисунке трансформации третьего уровня. Таким образом, оптимальным является правило «введение гидроксила в ароматическое кольца приводит к росту активности», основанное на закономерности с наивысшей статистической значимостью (на втором уровне иерархии слева на Рис. 15).

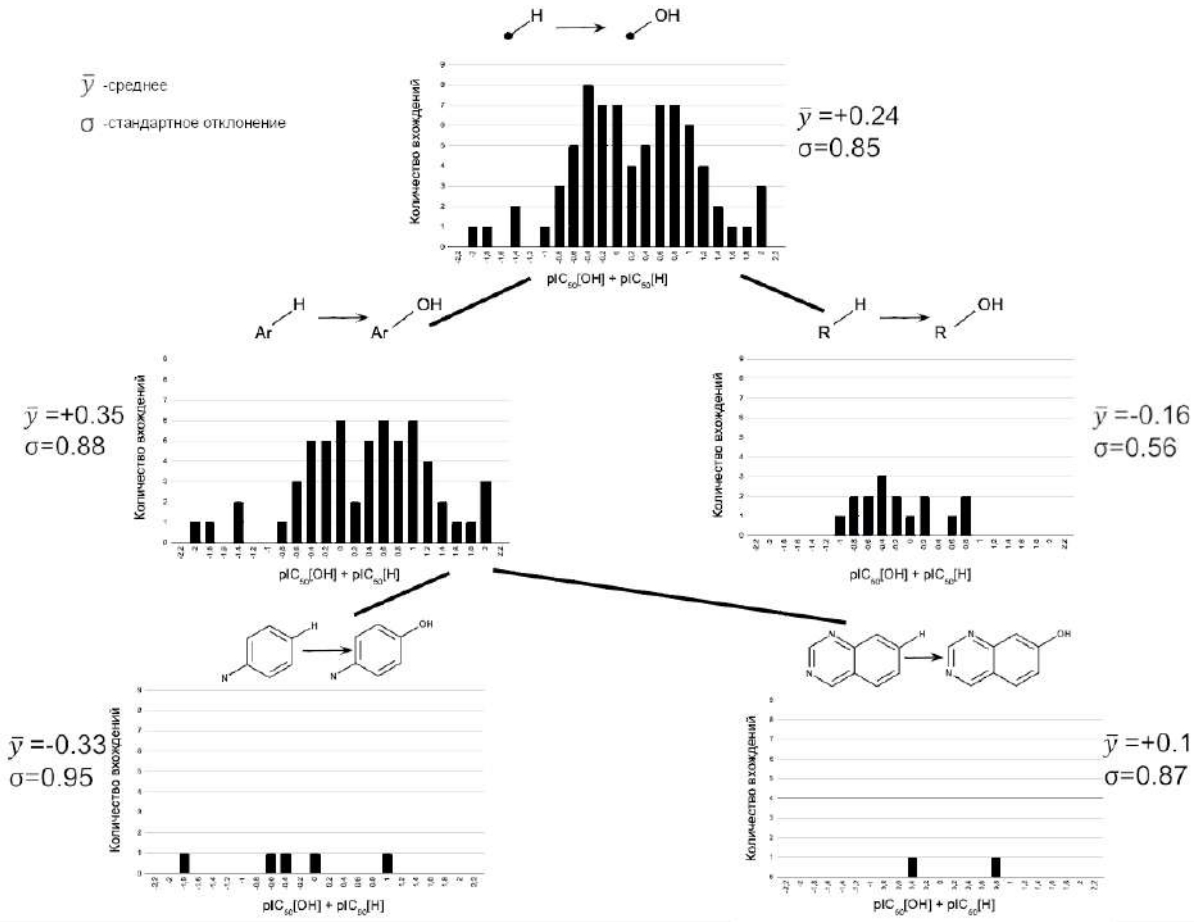


Рис. 15. Иерархия структурных контекстов для трансформации замены атома водорода на гидроксильную группу. $\bar{y}$ – средние значения, σ – стандартное отклонение, SEM. Рисунок адаптирован с публикации [54] и приводится с разрешения издательства. Copyright (2011) American Chemical Society.

2.1.4.4. Методы идентификации пар соответствия молекул в базах данных

К настоящему времени разработано большое число разнообразных алгоритмов идентификации пар соответствия молекул путем анализа содержимого баз данных. Все они могут быть отнесены к двум основным категориям. Методы «с учителем» (англ. *supervised*) используют заранее заданную трансформацию для извлечения соответствующих им пар ММР из базы данных. Для этого можно, например, путем подструктурного поиска найти молекулы, содержащие левую часть уравнения трансформации, применить к ней трансформацию всеми возможными способами и определить, содержатся ли в базе данных получающиеся при этом молекулы. Такой подход сводит задачу извлечения пар ММР к набору поисков в базе

данных и поэтому является эффективным в вычислительном плане. Тем не менее, необходимость заранее задавать трансформации делает его совершенно непригодным для работы в большинстве практически важных случаев, когда такой набор неизвестен и не может быть легко сформирован.

Наибольшей популярностью пользуются подходы из категории «без учителя» (англ. unsupervised), где сами трансформации также извлекаются из базы данных. Предложенные для этой цели подходы можно отнести к двум основным типам: на основе поиска наибольшей общей подструктуры (англ. Maximum Common Substructure, *MCS*) и на основе фиксированных схем фрагментации.

Методы на основе MCS базируются на алгоритмах поиска общего подграфа (пересечения графов), см. раздел 2.3.3 в пособии 2. Если для пары сходных молекул число атомов в найденной наибольшей общей подструктуре лишь немного меньше числа атомов в каждой из них, то левая часть трансформации может быть получена путем удаления общей подструктуры из первой молекулы, а правая часть – из второй. Хотя методы на основе MCS являются наиболее общими и найденные с их помощью решения наиболее корректными с точки зрения определения пар MMP (поскольку они рассматривают наименьшие изменения структуры при переходе от одной молекулы к другой), однако их слабой стороной является высокая вычислительная сложность, поскольку приходится решать задачу нахождения наибольшей общей подструктуры для всех пар молекул в базе данных. Тем не менее, были предложены приемы, значительно облегчающие решение этой задачи. В частности, один из наиболее эффективных из них сводится к проведению предварительной кластеризации базы данных с использованием мер сходства, определяемых на вычисляемых для молекул векторах дескрипторов либо фингерпринтах, что дает возможность проводить попарное пересечение химических структур только внутри кластеров небольшого размера. Подобные приемы, особенно взятые в сочетании с использованием высоко-параллельных вычислений, дают возможность извлекать пары MMP из баз данных очень большого размера (миллионы структур) [60-62].

В отличие от вышеупомянутых подходов, основанных на поиске наибольших общих подструктур, требующих очень больших вычислительных мощностей, подходы на основе фиксированных схем фрагментации очень эффективны в вычислительном плане, хотя и ограничены заданными схемами фрагментации. Вычислительная эффективность этой группы методов идентификации пар MMP

обусловлена тем, что каждая структура в базе данных обрабатывается только один раз, и поэтому вычислительная сложность возрастает пропорционально числу структур в базе (O(N)), тогда как в подходах на основе поиска наибольших общих подструктур рассматриваются пары структур, и поэтому вычислительная сложность возрастает пропорционально квадрату числа структур в базе (O(N^2)).

В рассматриваемых подходах фрагментация обычно ведется по ациклическим связям, т.е. по связям, не входящих в какие-либо циклы в молекуле. В работе [63] был предложен алгоритм идентификации пар ММР, основанный на фрагментации по всем имеющимся в молекуле ациклическим связям. На Рис. 16 приведена схема этого алгоритма в применении к парам ММР с одной точкой присоединения. На первом этапе ведется рассечение всех молекул из базы данных по всем ациклическим связям. Образовавшиеся в результате разрыва связей пары фрагментов организуют в словарь «ключ-значение». Далее для всех ключей, для которых было найдено таким образом несколько значений, из каждой пары значений для одного ключа образуют пару ММР.

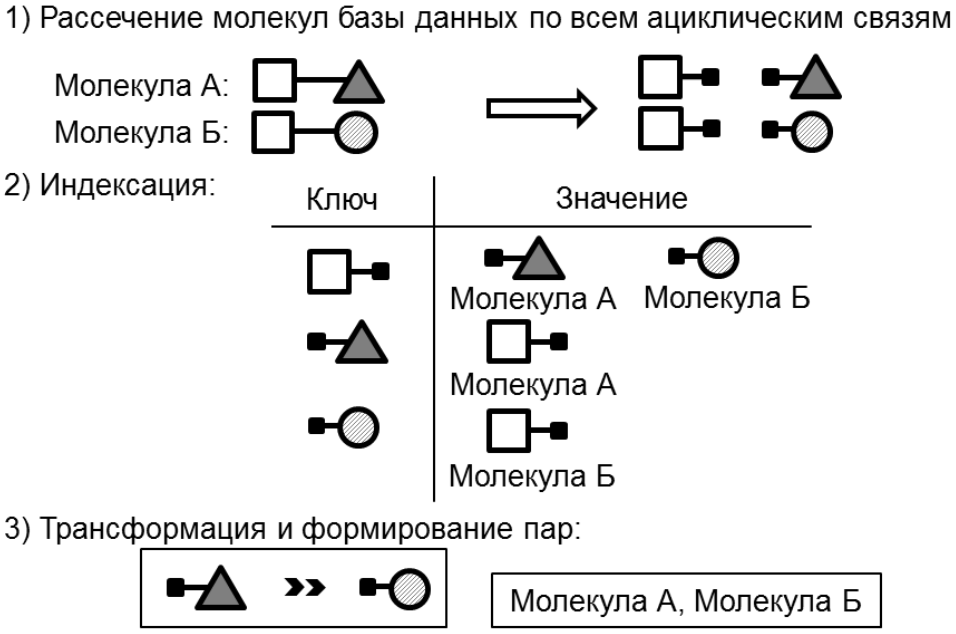


Рис. 16. Схема алгоритма идентификации пар ММР с одной точкой присоединения. Рисунок адаптирован из публикации [63] и приводится с разрешения издательства. Copyright (2010) American Chemical Society.

На Рис. 17 приведена схема этого же алгоритма в применении к парам ММР с двумя точками присоединения. В данном случае ведется рассечение молекул по всем возможным парам ациклических связей. После этого составляют словарь, ключи которого формируют из пар

образовавшиеся после разрыва связей фрагментов с одной точкой присоединения, а значения – из фрагментов с двумя точками присоединения. Далее, как и в предыдущем случае, для всех ключей, для которых было найдено несколько значений, из каждой пары значений для одного и того же ключа формируют пару ММР.

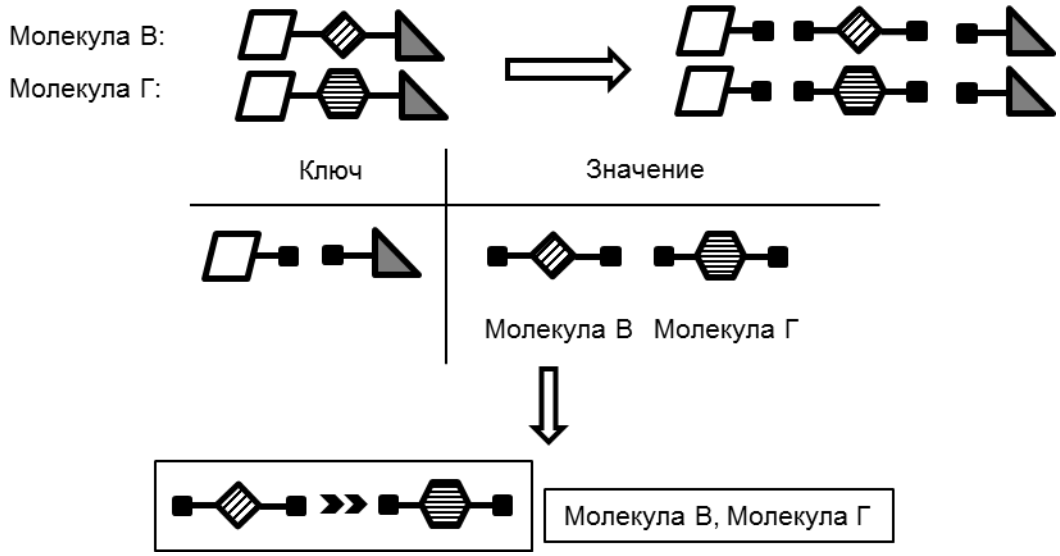


Рис. 17. Схема алгоритма идентификации пар ММР с двумя точками присоединения. Рисунок адаптирован из публикации [63] и приводится с разрешения издательства. Copyright (2010) American Chemical Society.

Одним из недостатков приведенного выше метода идентификации пар ММР является то, что рассматривается рассечение молекул по всем ациклическим связям, в результате чего формируется большое число пар молекул, трансформация между которыми весьма затруднительна в синтетическом плане. В работе [64] для решения этой проблемы было предложено рассматривать только легко реализуемые в органическом синтезе фрагментации, используя набор правил RECAP (*Retrosynthetic Combinatorial Analysis Procedure*) [65], представленный на Рис. 18. Получающиеся при этом «синтетически реализуемые» пары получили обозначение RECAP-ММР.

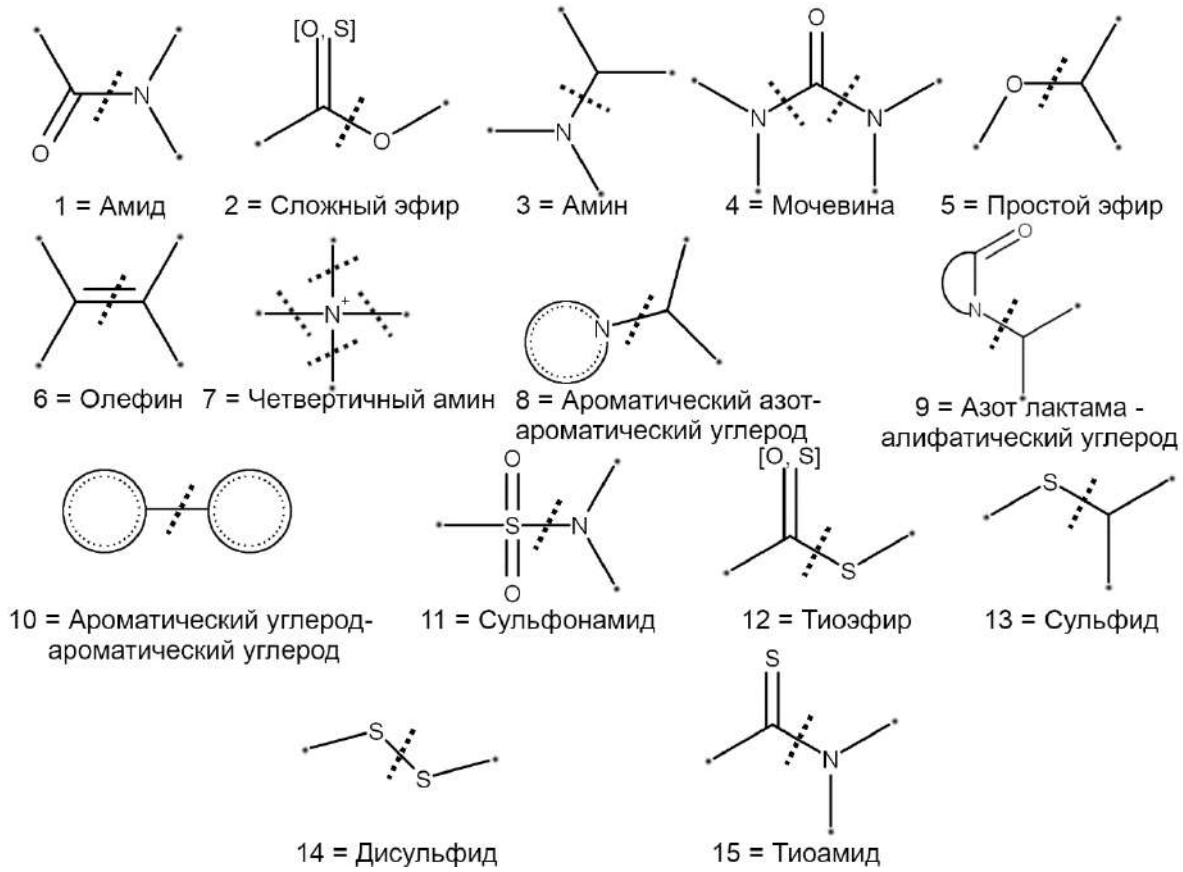
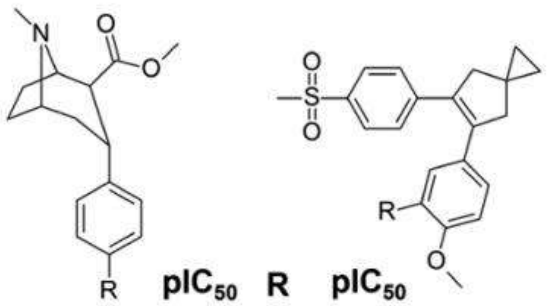


Рис. 18. Классические правила ретросинтетической фрагментации RECAP [64].

2.1.4.5. Ряды соответствия молекул

Естественным развитием представлений о парах соответствия молекул явилось формирование понятия о *рядах соответствия молекул* (англ. *Matched Molecular Series*, *MMS*), которые определяются как **наборы из двух или большего числа молекул с одним общим скелетом, но с разными заместителями в одном и том же положении** [66]. На Рис. 19 приведены в качестве примера два разных ряда соответствия молекул для одного и того же набора заместителей [H, F, Cl, Br]. Количественные данные по биологической активности для одного из них (слева) описывают связывание с транспортером дофамина, а для другого (справа) – ингибирование фермента COX-2 (циклооксигеназы-2). Ряды предпочтения (по понижению активности) для них разные – [Br > Cl > F > H] и [Br > H > F > Cl], соответственно.



pIC_{50}	R	pIC_{50}
7.00	H	8.30
7.68	F	8.00
8.51	Cl	7.77
8.77	Br	8.89
[Br>Cl>F>H]		**[Br>H>F>Cl]**

Рис. 19. Примеры рядов соответствия молекул для одного и того же набора заместителей, но рассматриваемых для разных остовов и разных свойств: связывания с транспортером дофамина (слева) и ингибирования фермента COX-2 (справа). Рисунок из публикации [66] приводится с разрешения издательства. Copyright (2014) American Chemical Society.

Целью анализа рядов соответствия молекул является оптимизация их свойств путем предсказания того, какие надо для этого ввести заместители в рассматриваемые ряды. В работе [66] для этого был предложен алгоритм Matsy (*MATched SEries*), использующий статистический анализ для того, чтобы предсказать, какой заместитель вероятнее всего приведет к дальнейшему повышению активности в данном ряду молекул для заданного порядка предпочтения. Рассмотрим алгоритм Matsy на конкретном примере, представленном на Рис. 20. Пусть в изучаемом ряду порядок предпочтения [A > B], что означает, что соединение с заместителем A более активно, чем соединение с заместителем B. Найдем в базе данных все ряды соответствия молекул, в которых A > B. На Рис. 20 их показано 5. Далее для каждого заместителя подсчитаем, сколько раз он встречается в найденных рядах (N) и в скольких при этом случаях активность соответствующего соединения превышает активность для A и, разумеется, B (N_+). Следовательно, в N_+ случаев его введение приводит к росту активности по сравнению с A, а в $N_- = N - N_+$ не приводит к росту активности. Далее, используя приведенные выше формулы (22) и (23) для биноминального распределения, находим p-значения для каждого из заместителей. Введение заместителя с минимальным p-

значением имеет наибольший шанс привести к повышению активности внутри рассматриваемого ряда соединений.

R Группа	N	N_+	$\frac{N_+}{N}$, %
D	3	3	100
E	1	1	100
C	4	1	25

Рис. 20. Схема алгоритма Matsy. Рисунок из публикации [66] приводится с разрешения издательства. Copyright (2014) American Chemical Society.

В работе [67] было предложено несколько количественных критериев для оценки сходства зависимостей «структура-активность» между рядами соответствия молекул, содержащих один и тот же набор заместителей. Высокий уровень сходства позволяет переносить известную зависимость с одного ряда на другой, что, в свою очередь, позволяет осуществлять прогноз активности в одном ряду соединения, пользуясь данными из другого. Было показано в этой же работе, что из рассмотренных вариантов к наилучшему результату приводит применение в качестве критерия сходства характеристики cRMSD (центрированного RMSD), вычисляемой по формуле:

$$cRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x}) - (y_i - \bar{y})]^2} \qquad (24)$$

где: n – число заместителей в ряду; x_i – активность i-го соединения в ряду x; $\bar{x}$ – среднее значение активности в ряду x; y_i – активность i-го соединения в ряду y; $\bar{y}$ – среднее значение активности в ряду y. Из нескольких вариантов осуществления прогноза был указан как наилучший вариант, вычисляемый по формуле:

$$p_{x-s,pred} = ap_{y-s} + b \qquad (25)$$

где $p_{x\text{-}s,pred}$ – предсказанная активность соединения в ряду *x* с заместителем *s*; $p_{y\text{-}s}$ – известная активность соединения в ряду *y* с этим же заместителем *s*; *a* и *b* – коэффициенты линейной регрессии, связывающей активности в рядах *x* и *y* у соединений с одинаковыми заместителями. В качестве ряда *y* выбирают ряд с наибольшим сходством с рядом *x*, для которого осуществляется прогноз.

2.2. ХИМИЧЕСКОЕ ПРОСТРАНСТВО ДЕСКРИПТОРОВ

Химическое пространство дескрипторов представляет собой многомерное векторное пространство, в котором молекулы представлены векторами дескрипторов. Имеется три основных подхода к визуализации и навигации в этом пространстве. Первый из них основан на принципе понижения размерности, в основе второго лежит кластеризация, а третий основан на построении графов соседства.

Стандартным методом понижения размерности является анализ главных компонент (PCA), см. раздел 2.3.2 в пособии 4. В рамках этого подхода несколько векторов (называемых *главными компонентами*), идущих вдоль главных осей инерции облака точек в пространстве дескрипторов, используются как базис нового низкоразмерного пространства, в которое проецируются точки из исходного пространства дескрипторов. При этом проецировании происходит минимальная потеря информации и, следовательно, максимально возможное сохранение отношение соседства между точками. Благодаря этому такую проекцию в низкоразмерное пространство можно рассматривать как «навигационную карту» в химическом пространстве дескрипторов. На этой идее основана работа систем ChemGPS (*chemical global positioning system*) [68] и ChemGPS-NP [69, 70], которые осуществляют *глобальное позиционирование* химических соединений в пространстве «лекарствоподобных» структур. Глобальное позиционирование основано на использовании в качестве координат главных компонент, полученных в рамках «универсальной» модели PCA, построенной для одного специально выбранного набора химических структур и молекулярных дескрипторов для их описания. В отличие от глобального, *локальное позиционирование* основано на использовании «локальных» моделей PCA, которые строятся отдельно для разных наборов химических структур и молекулярных дескрипторов и поэтому не являются универсальными. Таким образом, глобальное позиционирование однозначно задает положение химического соединения в «универсальном» химическом

пространстве дескрипторов таким же образом, как система глобального позиционирования GPS однозначно задает положение любого объекта на поверхности Земли. В то же время модели для глобального позиционирования химических соединений построены таким образом, чтобы отражать те отношения соседства, которые выявляются при помощи локальных моделей. В отличие от карт, построенных с помощью локальных моделей, карты, построенные с помощью глобального позиционирования, корректно сравнивать между собой. В качестве примера на Рис. 21 приведены карты, построенные для набора из 8599 монокарбоновых кислот с помощью глобального позиционирования в рамках системы ChemGPS (слева) и с помощью локальной модели, построенной с помощью метода PCA (справа). На рисунке также приведены метки для нескольких химических соединений. Легко видеть, что карты показывают фактически одну и ту же картину соседства, что свидетельствует о возможности использовать «универсальные» карты, основанные на глобальном позиционировании, вместо множества локальных.

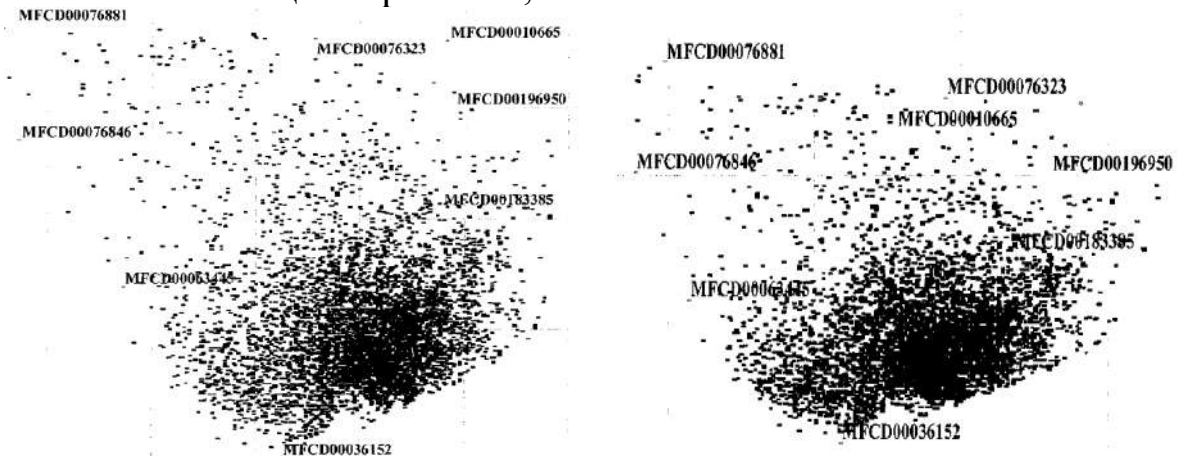


Рис. 21. (слева) Карта для 8599 монокарбоновых кислот, построенная при помощи системы глобального позиционирования ChemGPS. (справа) Карта, построенная с помощью одной из локальных моделей PCA. Рисунок из публикации [68] приводится с разрешения издательства. Copyright (2001) American Chemical Society.

Хотя оси «навигационной карты» PCA ортогональны, соответствующие латентные переменные являются статистически независимы только тогда, когда они подчиняются Гауссовому (нормальному) распределению. Поскольку же в реальности распределение значений молекулярных дескрипторов далеко от нормального, то и латентные переменные становятся статистически взаимосвязанными, что существенно затрудняет их

интерпретируемость и мешает практическому применению построенным на их основе «картам навигации». Для решения этой проблемы был предложен метод независимых компонент (англ. *Independent Component Analysis, ICA*) [71-74], который приводит к формированию статистически независимых латентных переменных, см. раздел 2.14.2.1 пособия 4. Было показано, что использование ICA вместо PCA при обработке химических данных приводит к формированию более легко интерпретируемых латентных переменных [75].

Альтернативный подход к картированию и навигации по химическому пространству дескрипторов основан на использовании иерархического кластерного анализа (см. раздел 2.14.1.2 в пособии 4). Получающиеся в результате такого анализа дендрограмы содержат богатую информацию о соотношениях соседства между объектами в химическом пространстве, а также между их кластерами. Ясность картины при этом, однако, исчезает при переходе к большому числу соединений. Тем не менее, в литературе описаны случаи применения иерархического кластерного анализа для анализа больших баз химических данных. В частности, Аграфиотис с соавт. [76] предложили для этого использовать радиальные кластерограммы, разные сегменты которых можно окрасить в соответствии с биологической активностью или любыми другими свойствами химических соединений.

Набор химических соединений также может быть описан как граф, в котором вершины соответствуют индивидуальным соединениям, а ребра соединяют сходные (в соответствии с выбранной мерой сходства) соединения [77]. Этот подход, в частности, был использован для описания взаимосвязи между разными классами биологически активных молекул [78], для описания сходства внутри наборов биологически активных соединений [79], а также для описания зависимости селективности от химической структуры [80].

2.2.1. Описание химического пространства дескриптора при помощи самоорганизующихся карт Кохонена (SOM)

Самоорганизующиеся карты Кохонена (SOM) подробно рассмотрены в разделе 2.11.4 пособия 4. Они являются в настоящее время, наряду с PCA, одним из самых популярных способов понижения размерности данных и визуализации химического пространства. В отличие от PCA, карты Кохонена осуществляют нелинейную проекцию данных, что может привести к лучшему их

сжатию. Кроме того, карты Кохонена являются «квантователями данных», то есть разбивают данные на группы схожих объектов, вследствие чего обучающая выборка имеет тенденцию быть равномерно распределенной по всей карте. Это приводит к высокой информационной насыщенности карты и возможности увидеть тонкие детали, характеризующие распределение данных. Кроме того, карты Кохонена обладают способностью сохранять отношение соседства: близкие структуры будут отображены либо в одну, либо в соседние ячейки. Эта особенность SOM имеет далеко идущие последствия. Поскольку близкие структуры имеют тенденцию обладать сходными свойствами, то химические соединения со сходными свойствами (например, действующие на одну и ту же биологическую мишень) имеют тенденцию отображаться в одну и ту же, либо в соседние ячейки SOM. Вследствие этого, классы активности оказываются распределенными по карте крайне неравномерно, и вся карта Кохонена оказывается разбитой на протяженные области, в каждой из которых преобладают соединения определенного класса активности. В определенном смысле SOM можно уподобить политической карте мира, которая разбита на страны, в каждой из которых обычно преобладает один народ. Благодаря этому карты Кохонена часто обладают неплохой прогнозирующей способностью, поскольку по тому, в какую область попадет химическое соединение, можно предсказать, какому классу активности оно принадлежит. То, насколько хорошо SOM распределяет соединения разных классов активности по разным областям, является важнейшим критерием качества SOM.

При рассмотрении методологии работы следует различать карты SOM низкого и высокого разрешения. В SOM низкого разрешения число ячеек (нейронов) значительно меньше, чем число химических соединений в обучающей выборке. Соответственно, в каждую ячейку попадает множество соединений. В качестве примера работы с картами низкого разрешения рассмотрим приведенную на Рис. 22 карту SOM размера 6x8, построенную в работе [81] на выборке, состоящей из субстратов Р-гликопротеина, ингибиторов и веществ с двойным действием, с использованием в качестве дескрипторов топологических индексов и дескрипторов электротопологического состояния. При визуализации карты использована т.н. U-матрица Ултша [82], представляющая собой матрицу евклидовых расстояний между векторами весов соседних ячеек (нейронов). Значения элементов U-матрицы показаны на рисунке при помощи шкалы серого цвета [83]. Чем светлее цвет между ячейками, тем ближе в химическом

пространстве дескрипторов они расположены. Красным цветом изображены ячейки, в которые попали только субстраты Р-гликопротеина (размер ячеек пропорционален их числу) из обучающей выборки, в зеленые ячейки попали только ингибиторы Р-гликопротеина, в синие – соединения с двойным действием, а в ячейки с несколькими цветами – соединения, относящиеся к нескольким классам биологической активности.

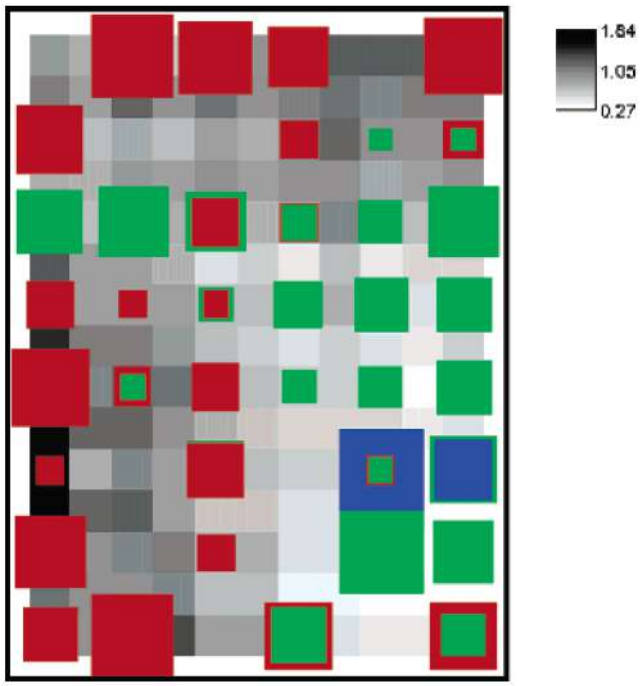


Рис. 22. Визуализация при помощи 6x8 SOM низкого разрешения с использованием U-матрицы субстратов и ингибиторов Р-гликопротеина. Красный, зеленый и синий цвет соответствуют, соответственно, субстратам Р-гликопротеина, его ингибиторам и соединениям с двойным действием. Оттенки серого цвета кодируют значения элементов U-матрицы. Рисунок из публикации [81] приводится с разрешения издательства. Copyright (2005) American Chemical Society.

Как видно на Рис. 22, соединения разных классов активности занимают на карте отдельные области. Также можно заметить, что субстраты разбросаны по большой площади, что следует из темной окраски соответствующих областей, указывающей на большие расстояния между ними в исходном дескрипторном пространстве. Это свидетельствует о широкой субстратной специфичности Р-гликопротеина. В отличие от них ингибиторы сконцентрированы в компактной области (что следует из светлой окраски, задаваемой элементами U-матрицы). Это свидетельствует о наличии более

жестких условий, необходимых для проявления ингибиторной активности по отношению к этому белку.

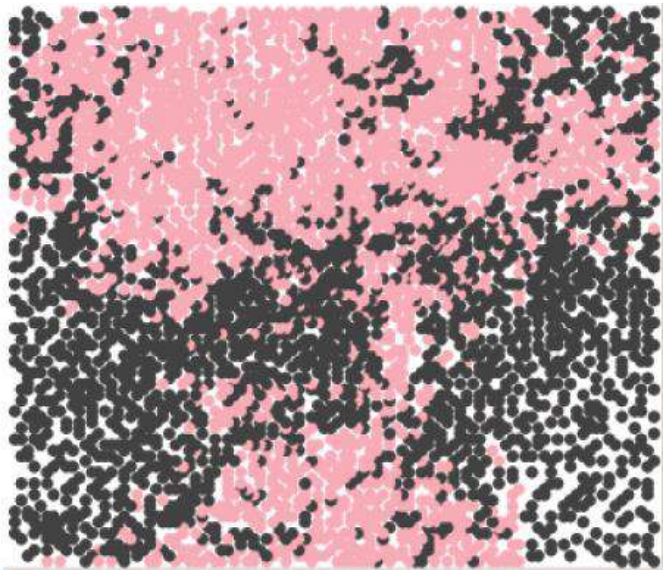


Рис. 23. Карта SOM высокого разрешения для 2653 лигандов GPCR-рецепторов (розовый цвет) и 2726 соединений, не являющимися лигандами GPCR-рецепторов (черный цвет). Рисунок из публикации [84] приводится с разрешения издательства. Copyright (2004) American Chemical Society.

В SOM высокого разрешения число ячеек (нейронов) карты значительно меньше числа химических соединений в обучающей выборке, и поэтому не все ячейки заселены, и каждая из заселенных содержит, как правило, не больше одного соединения. Рассмотрим в качестве примера приведенную на Рис. 23 карту SOM высокого разрешения, построенную в работе [84] на выборке, содержащей соединения двух классов: (1) лигандов GPCR-рецепторов (розовый цвет); (2) соединений, не являющимися лигандами GPCR-рецепторов (черный цвет). Размер карты - 100x100 ячеек, организованных в виде тора. Для представления химических структур были использованы топологические дескрипторы на основе пар фармакофорных центров. На рисунке видно явное преобладание соединений каждого из двух классов в разных областях карты. Следовательно, по тому, в какую из этих областей будет отображено тестовое соединение, можно предсказать, будет ли оно связываться с каким-либо из GPCR-рецепторов.

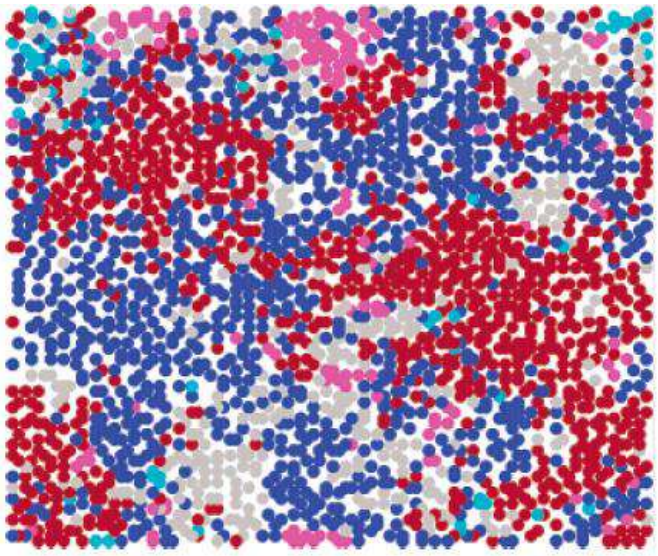


Рис. 24. Карта SOM высокого разрешения для лигандов GPCR-рецепторов. Красным цветом обозначены аминергические рецепторы семейства I, синим – пептидергические рецепторы семейства I, серым – остальные типы рецепторов семейства I, пурпурным – рецепторы семейства II, голубым – рецепторы семейства III. Рисунок взят из публикации [84] и приводится с разрешения издательства. Copyright (2005) American Chemical Society.

На Рис. 24 приведена взятая из той же самой публикации карта SOM, построенная только на лигандах GPCR-рецепторов, относящихся к 5 классам активности: (1) лиганды аминергических рецепторов семейства I; (2) пептидергических рецепторов семейства I; (3) остальных рецепторов семейства I; (4) рецепторов семейства II; (5) рецепторов семейства III. Таким образом, соединения, занимающие на Рис. 23 область розового цвета, оказались равномерно распределенными по всей карте на Рис. 24. Как и в предыдущем случае, наблюдается четкая тенденция распределения лигандов, относящихся в данном случае к разным семействам GPCR-рецепторов, по разным областям карты.

Рис. 25 демонстрирует кластеризующую способность карт Кохонена. На нем приведена та же самая карта SOM, на которой желтым цветом выделены лиганды четырех GPCR-рецепторов: (1) аденозинового A_{2A} (аминергический рецептор семейства I), (2) каннабиноидного (CB, рецептор семейства I, не являющийся ни аминегическим ни пептидергическим), (3) эндотелинового (ET, пептидергический рецептор семейства I), (4) CRF (рецептор семейства II). Поскольку аденозиновый A_{2A} рецептор (Рис. 25a) не имеет подтипов, то его лиганды на карте SOM образуют один выраженный

кластер. Для CB (Рис. 25b) и CRF (Рис. 25c) рецепторов на рисунке видно по два основных кластера, которые соответствуют наличию двух подтипов у каждого их этих рецепторов. На Рис. 25d, однако, виден один ярко выраженный кластер, несмотря на наличие двух подтипов у ET-рецептора, что можно объяснить большим сходством их сайтов связывания и, как следствие, строения лигандов.

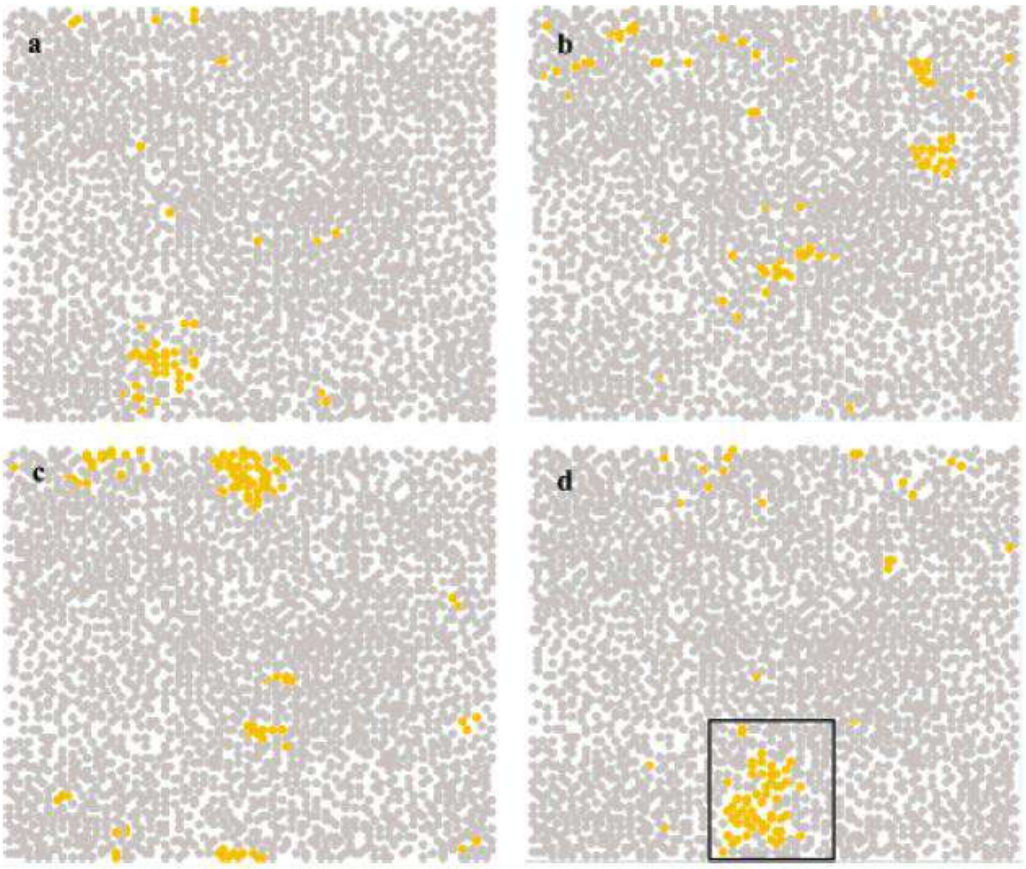


Рис. 25. Карта SOM высокого разрешения для лигандов GPCR-рецепторов. Желтым цветом выделены лиганды следующих рецепторов: (a) аденозинового A_{2A}, (b) каннабиноидного (CB), (c) CRF, (d) эндотелинового (ET). Рисунок из публикации [84] приводится с разрешения издательства. Copyright (2004) American Chemical Society.

Следует, однако, отметить, что попадание молекул в один кластер на карте SOM не всегда означает, что они относятся к одному структурному классу. Например, при использовании дескрипторов, построенных на основе фармакофоров, как в рассмотренном выше случае, в один кластер могут попасть структурно разнородные соединения, но имеющие сходное расположение фармакофорных центров. Это явление называется «scaffold hopping». В качестве примера рассмотрим ограниченную прямоугольником область карты

SOM в нижней части Рис. 25d, содержащую лиганды эндотелинового рецептора. В увеличенном виде эта область представлена на Рис. 26, в правой части которого представлены структурные формулы некоторых из присутствующих там лигандов эндотелинового рецептора. Видно, что они являются представителями разных структурных классов, хотя и элементы сходства тоже отчетливо видны. Из этого следует, что данной картой можно пользоваться при виртуальном скрининге с целью выявления лигандов определенных типов (например, эндотелинового) GPCR-рецепторов, причем благодаря «scaffold hopping» есть возможность выявить таким образом лиганды, относящиеся к новым структурным классам и, следовательно, незапатентованные.

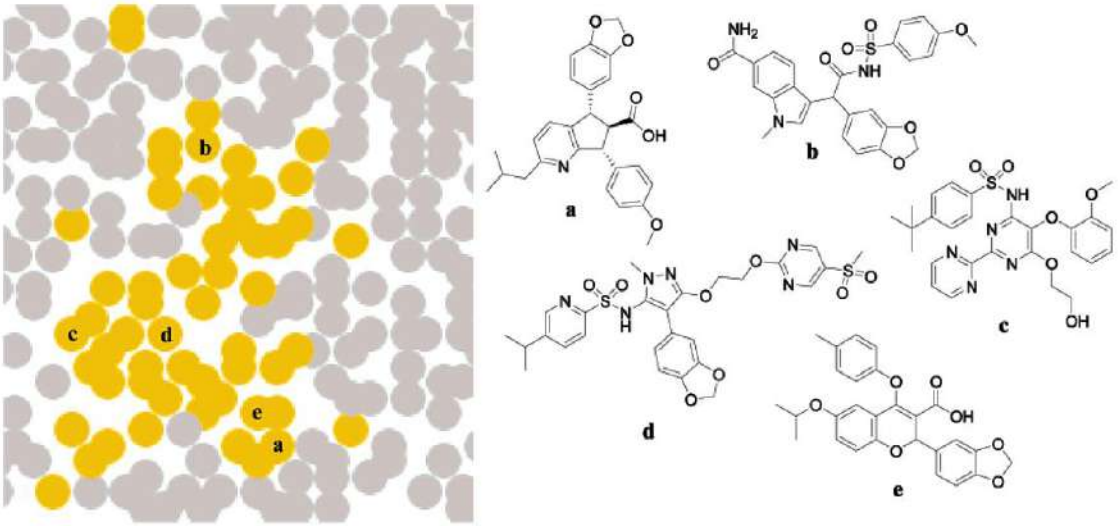


Рис. 26. Часть карты лигандов GPCR-рецепторов, содержащие кластер лигандов эндотелинового рецептора. Рисунок из публикации [84] приводится с разрешения издательства. Copyright (2004) American Chemical Society.

2.2.2. Описание химического пространства дескрипторов при помощи генеративных топографических отображений (GTM)

Метод генеративных топографических отображений GTM подробно описан в разделе 2.13 пособия 4. Его можно считать альтернативой использованию метода главных компонент PCA и самоорганизующихся карт Кохонена (SOM) для построения карт химического пространства дескрипторов. Рассмотрим преимущества, которые дает использование GTM для этой цели.

2.2.2.1. Сравнение карт, построенных с помощью GTM, PCA и SOM

На Рис. 27 приведена визуализации выборки, состоящей из лигандов, связывающихся с 10 разными биомишенями (*ache* (ацетилхолинэстераза), *cox2* (циклооксигеназа-2), *dhfr* (дигидрофолатредуктаза), *egfr* (эпидермальный фактор роста), *fgfr1* (киназа рецепторов фактора роста фибробластов), *fxa* (фактор Xa), *p38* (митоген активирующий белок), *pdgfrb* (киназа рецепторов фактора роста производных тромбоцитов), *src* (тирозинкиназа), *vegfr2* (рецептор фактор роста эндотелия сосудов)), при помощи трех методов: GTM, PCA и SOM [85].

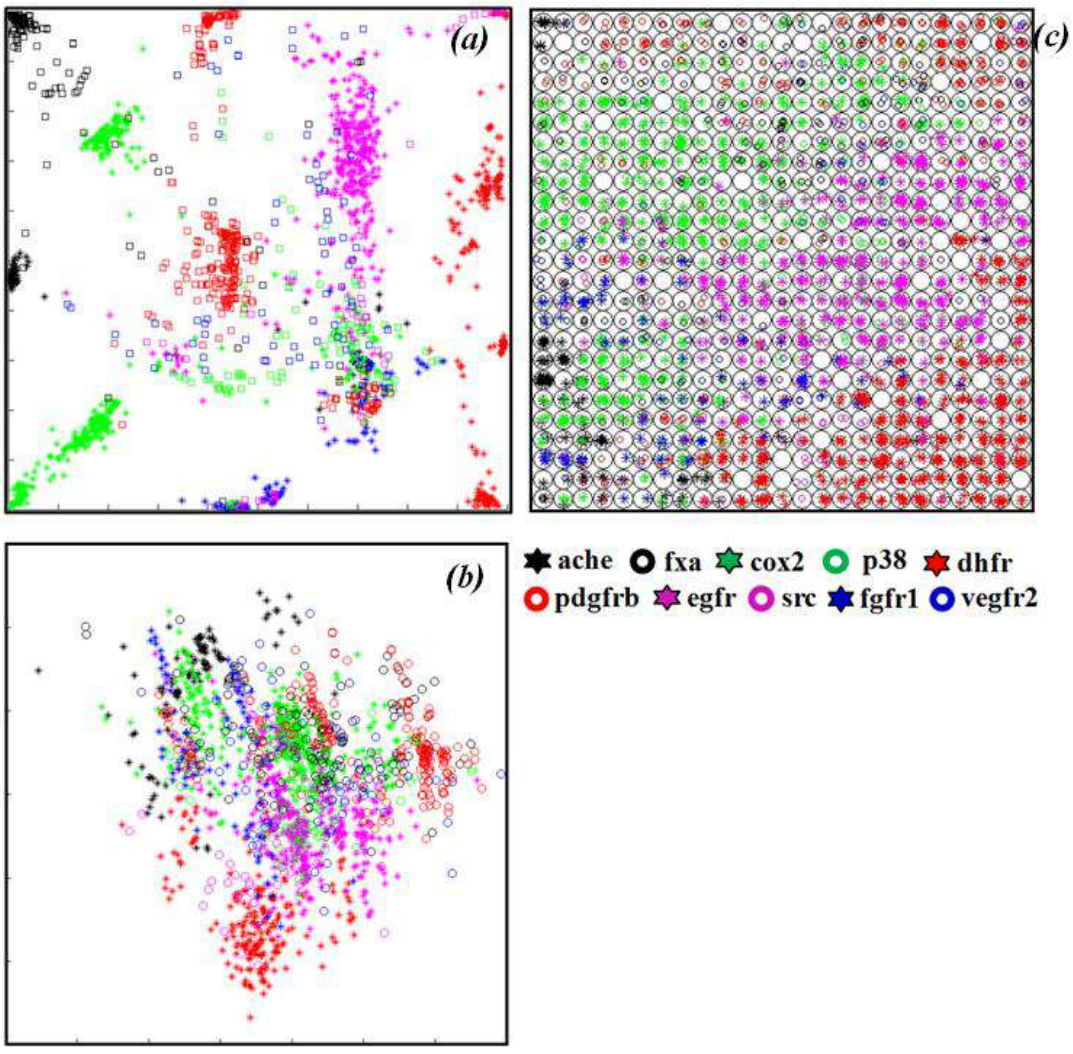


Рис. 27. Визуализации выборки лигандов 10 биомишеней при помощи GTM ***(a),*** PCA (***b***) и SOM (***c***). Рисунок из публикации [85] приводится с разрешения издательства. Copyright (2012) WILEY-VCH Verlag.

Преимущество GTM очевидно. На графике PCA соединения, представляющие разные классы активности, плохо отделены друг от друга и занимают относительно небольшую область в его центре. На графике SOM, в отличие от предыдущего случая, классы хорошо отделены друг от друга, однако, наглядно видны другие проблемы. В частности, практически все ячейки карты SOM оказываются занятыми, что должно привести к очевидно неправильному выводу о том, что и все остальные химические структуры должны быть лигандами хотя бы одной из 10 вышеупомянутых биомишеней.

2.2.2.2. Классификационные ландшафты активности

Наряду с регрессионными ландшафтами активности, рассмотренными в разделе 2.13.2 пособия 4, метод GTM позволяет также строить *классификационные ландшафты активности* (англ. *classification activity landscapes*), описывающие распределение соединений, относящихся к определенному классу активности, по карте [86]. Для этой цели для заданного класса активности *C* находят для каждого угла *k* решетки значения *кумулятивной ответственности* (англ. *cumulated responsibility*) ρ_{Ck} по формуле:

$$\rho_{Ck} = \sum_{n \in C} R_{kn} \qquad (26)$$

где R_{kn} – ответственность узла *k* за соединение *n* (см. раздел 2.13.1 в пособии 4), причем суммирование ведется по всем соединениям *n*, обладающим активностью класса *C*. В качестве эталона сравнения для класса *C* вычисляется характеристика ρ_C^0, равная среднему числу представителей класса *C*, приходящихся на один узел решетки:

$$\rho_C^0 = \frac{N_C}{K} \qquad (27)$$

где N_C – число представителей класса *C* в выборке, *K* – число узлов в решетке. Это дает возможность найти ρ_{Ck}^* - нормализованную плотностью класса *C* на узле *k* решетки:

$$\rho_{Ck}^* = \rho_{Ck} / \rho_C^0 \qquad (28)$$

Тогда можно считать, что узел *k* заселен преимущественно представителями класса *C*, если для всех альтернативных классов $C' \neq C$ справедливо неравенство $\rho_{Ck}^* > \rho_{C'k}^*$. В частности, при рассмотрении одного типа биологической активности все соединения можно условно считать принадлежащими двум классам – активных, обозначаемого как «2», и неактивных, «1». Это дает возможность для каждого узла *k* найти значение характеристики $\bar{C}_k$:

$$\bar{C}_k = \frac{2 \times \rho^*_{2k} + \rho^*_{1k}}{\rho^*_{2k} + \rho^*_{1k}} \quad (29)$$

В этом случае величины $\bar{C}_k$, близкие к 2, означают заселенность узла преимущественно активными соединениями, а близкие к 1 – неактивными.

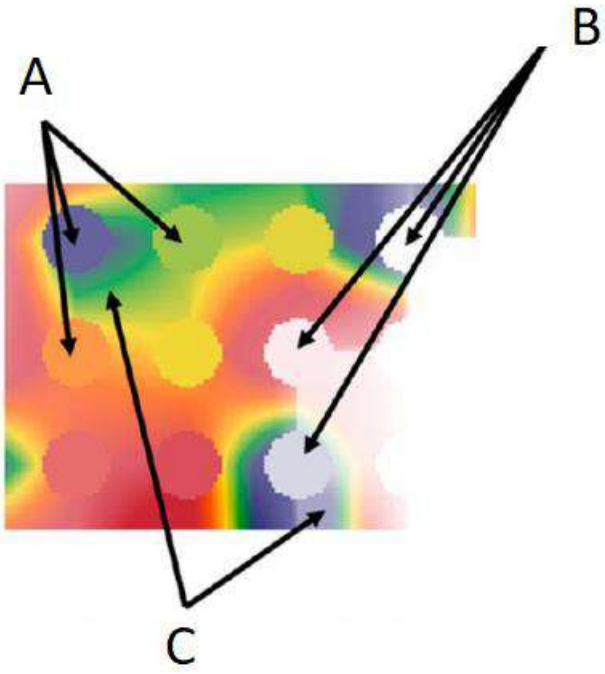


Рис. 28. Карта классификационного ландшафта активности. Применена кодировка распределения характеристики $\bar{C}_k$ пятью цветами: (1) красный цвет обозначает зоны преимущественно неактивных соединений ($\bar{C}_k < 1.4$); (2) оранжевый цвет обозначает зоны небольшого преобладания неактивных соединений ($1.4 \leq \bar{C}_k < 1.5$); (3) желтый цвет обозначает зоны небольшого преобладания активных соединений ($1.5 \leq \bar{C}_k < 1.6$); (4) зеленый цвет обозначает зоны явного преобладания активных соединений ($1.6 \leq \bar{C}_k < 1.7$); (5) синий цвет обозначает зоны преимущественного преобладания активных соединений ($\bar{C}_k \geq 1.7$). **A** – узлы решетки, обозначенные кругами, равномерно заполненными окраской с одинаковой интенсивностью; **B** – узлы с очень низкими значениями суммарной плотности, обозначенные окраской очень слабой интенсивности либо отсутствием окраски; **C** – промежутки между узлами окрашены в соответствии с цветом и интенсивностью, полученной интерполяцией по ближайшим узлам. Рисунок из публикации [86] приводится с разрешения издательства. Copyright (2016) American Chemical Society.

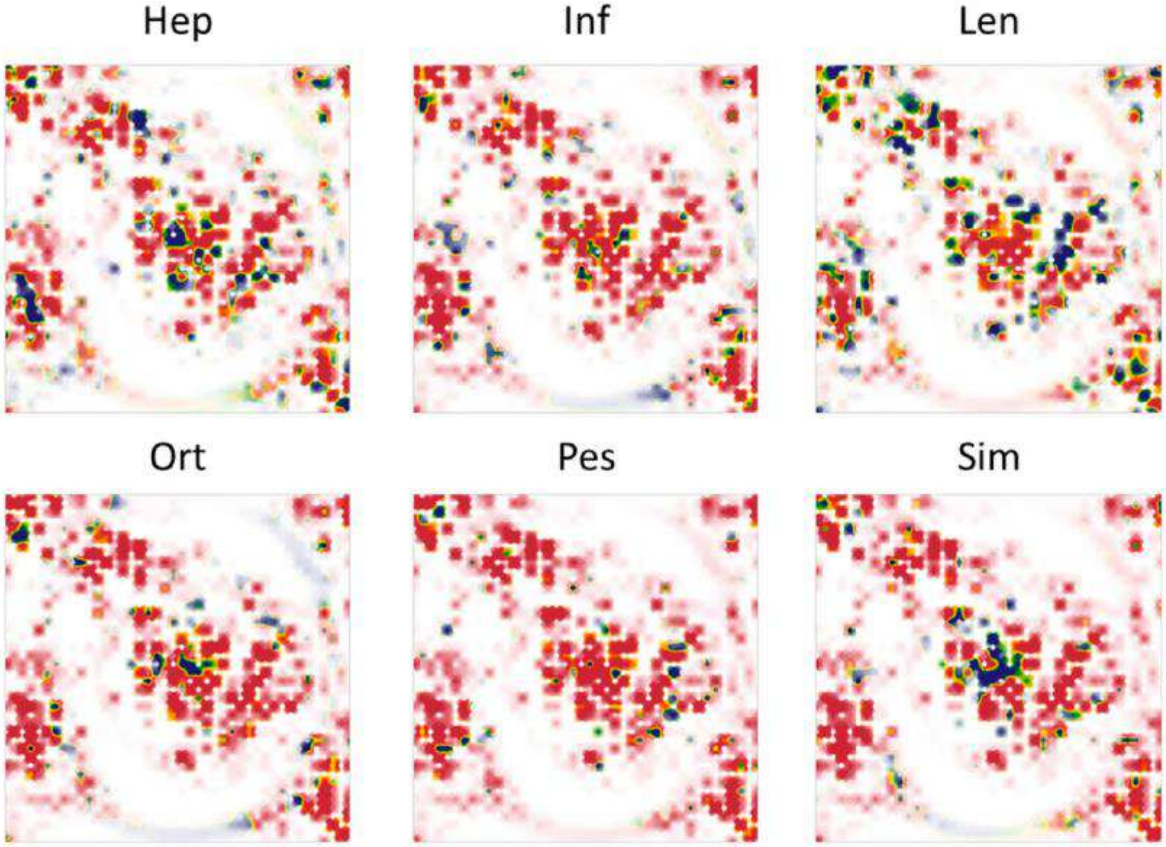


Рис. 29. Карты классификационного ландшафта для соединений с антивирусной активностью, действующих на 6 классов вирусов: Hep (вирус гепатита C), Inf (вирус гриппа A), Len (вирус иммунодефицита человека HIV), Ort (вирус гепатита B), Pes (вирус диареи крупного рогатого скота), Sim (вирус герпеса). Области активных соединений обозначены синим цветом. Рисунок из публикации [86] приводится с разрешения издательства. Copyright (2016) American Chemical Society.

Построение классификационных ландшафтов активности основано на кодировании на карте GTM величин $\bar{C}_k$ цветом, а суммарной плотности точек (в данном случае $\rho_{1k}+\rho_{2k}$) – интенсивностью закраски при превышении суммарной плотностью некоторого порога, либо отсутствием закраски в малозаселенных областях карты, где суммарная плотность меньше этого порога. В этом случае удобно обозначать узлы решетки кругами, равномерно окрашенными постоянным цветом с одинаковой интенсивностью, тогда как для закрашивания промежутков между узлами соответствующие цвета и интенсивности могут быть найдены интерполяцией по ближайшим узлам, Рис. 28. В качестве примера на Рис. 29 представлены карты классификационного ландшафта

активности, построенные в работе [86] для антивирусных препаратов, действующих на шесть классов вирусов.

2.2.2.3. Привилегированные шаблоны ответственности и привилегированные структурные мотивы

В идеальном случае для заданной молекулы *m* набор значений ответственности R_{km}, вычисленный для всех узлов *k*, должен однозначно характеризовать свойства этого химического соединения. Это значит, что две молекулы, обладающие одинаковыми значениями R_{km} для всех узлов решетки, являются полностью эквивалентными с точки зрения любого проводимого с помощью GTM анализа и поэтому в идеале должны обладать одинаковыми свойствами. Практическому применению этого принципа, однако, мешает то, что значения R_{km} являются действительными числами с непрерывным диапазоном значений, а точное равенство возможно только между числами с дискретным набором значений. Тогда можно ожидать, что кластеры, образованные молекулами с одинаковыми дискретными значениями ответственностей для всех узлов решетки, будут характеризоваться одинаковым профилем свойств (активности).

Привилегированный шаблон ответственности RP – это вектор значений RP_{km}, образованный значениями ответственности R_{km} после дискретизации. В частности, в работе [86] была предложена следующая схема перехода от R_{km} к RP_{km}: $RP_{km} = 0$ при $R_{km} < 0.01$; $RP_{km} = 1$ при $0.01 \leq R_{km} < 0.11$; $RP_{km} = 2$ при $0.11 \leq Rkm < 0.21$; и т.д. Тогда можно считать, что все соединения, обладающие одинаковым шаблоном ответственности, принадлежат одному кластеру структурно близких молекул, которые могут характеризоваться одинаковым набором свойств.

Привилегированный структурный мотив – это общий структурный мотив (например, общая подструктура либо соответствие общей структуре Маркуша), характеризующий набор молекул с одинаковым привилегированным шаблоном ответственности. Выявление таких мотив путем анализа карт GTM предоставляет в руки медицинского химика очень ценную информацию. В качестве примера на Рис. 30 приведены найденные в работе [86] путем анализа карт GTM привилегированные структурные мотивы антивирусных препаратов.

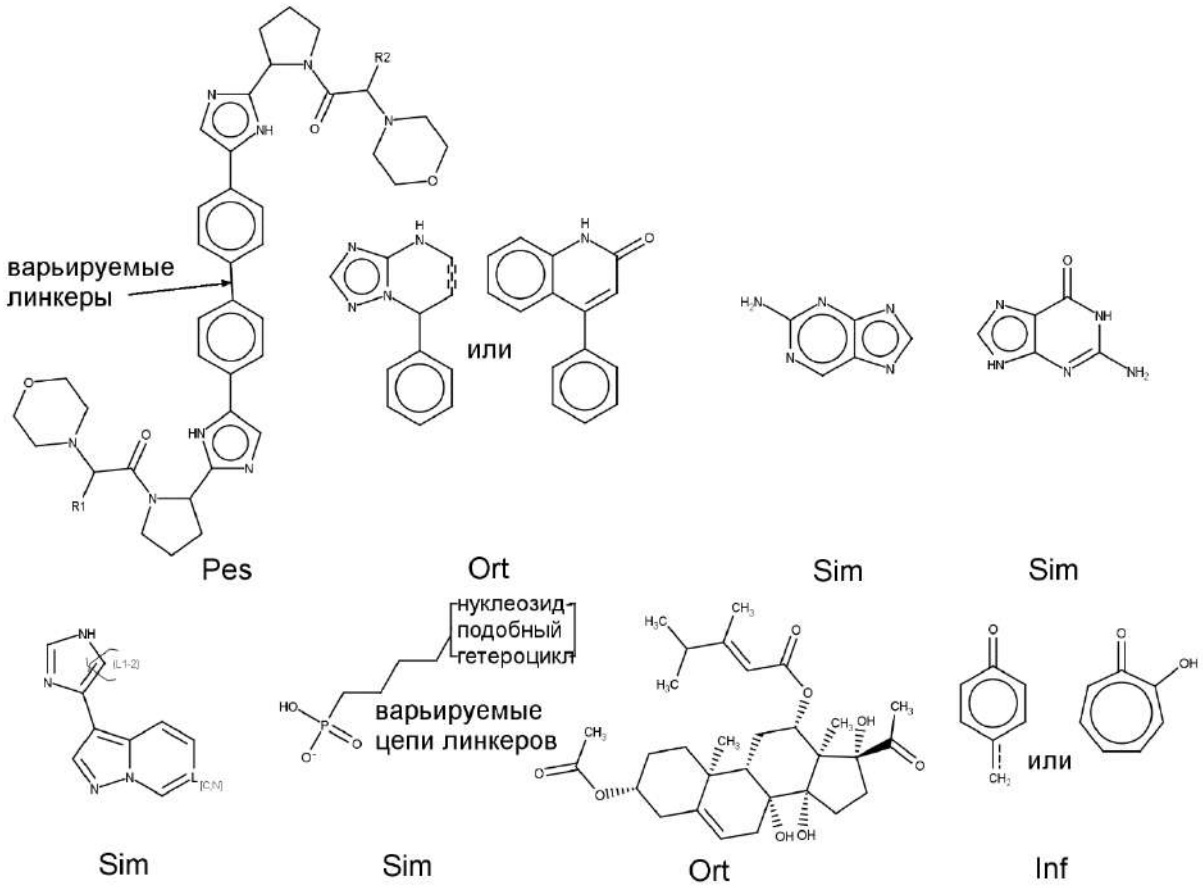


Рис. 30. Привилегированные структурные мотивы для антивирусных препаратов, действующих на вирусы, принадлежащие нескольким классам: Inf (вирус гриппа A), Ort (вирус гепатита B), Pes (вирус диареи крупного рогатого скота), Sim (вирус герпеса).

2.2.2.4. Иерархическая визуализация больших баз данных

При использовании GTM для визуализации и анализа химических баз данных очень большого размера, которые могут насчитывать миллионы и даже миллиарды соединений, возникает проблема «перегруженности» узлов карты, когда на узел решетки отображается очень большое число соединений, что очень затрудняет использование карты для анализа распределений активности и выявления привилегированных структурных мотивов (см. выше). Для решения этой проблемы было предложено визуализировать базы данных иерархическим образом [87-89]. В этом случае сначала строится карта для всей базы данных, затем пользователь выбирает в ней «интересную» область небольшого размера, после чего карта перестраивается с использованием только тех соединений, которые отображаются главным образом на выбранную область, обеспечивая

тем самым «увеличение» для выбранной области на карте. Эта операция может быть повторена и с построенной картой, и так до тех пор, пока на каждый узел будет отображаться относительно небольшое число соединений, позволяющее эффективно проводить анализ «структура-активность» и находить привилегированные структурные мотивы. Таким образом, иерархических принцип визуализации позволяет работать с базами данных практически неограниченного объема. На Рис. 31 приведена в качестве примера иерархическая визуализация большой базы, включающей 21 миллион соединений, что достигается построением дополнительных карт на подвыборках второго и третьего уровня, состоящих из, соответственно, 650 тысяч и 2500 соединений, которые позволяют сконцентрироваться на анализе выбираемых пользователем «интересных» областей.

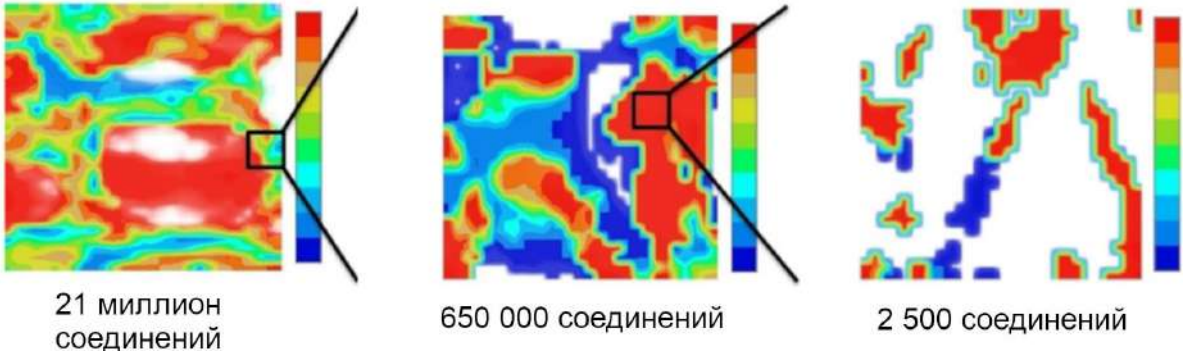


Рис. 31. Трехуровневая иерархическая визуализация базы данных, состоящей из 21 миллиона соединений. Пример из публикации [89] приводится с разрешения издательства. Copyright (2017) Wiley VCH.

2.2.2.5. Универсальные карты на основе GTM

Карты, построенные с помощью GTM, могут быть локальными и глобальными (универсальными), так же, как и в случае рассмотренных выше карт на основе PCA (см. в начале раздела 2.2). Локальные карты обеспечивают локальное позиционирование, зависящее от выбранной базы данных и набора используемых дескрипторов. При построении модели GTM для универсальной карты используется единый, специально выбранный набор данных и оптимальный (в соответствии с определенными критериями) набор дескрипторов. В этом случае строится универсальная карта GTM (манифолд), проекции на который как можно большего числа выборок должна приводить к построению ландшафтов активности, обладающих в среднем наивысшей прогнозирующей способностью. В частности, регрессионные ландшафты должны обеспечить в среднем наибольшую точность

прогнозирования, а классификационные – наилучшее разделение активных и неактивных молекул.

В работе [90] были построены универсальные модели GTM (манифолды) на основе данных по сотням видам биологической активности, извлеченных из ChEMBL. Оптимизация параметров модели и поиск оптимального набора дескрипторов осуществлялись с помощью генетического алгоритма, где в качестве критерия оптимизации выступал параметр, зависящий от усредненной прогнозирующей способности соответствующих ландшафтов по отношению к большому числу типов биологической активности. Благодаря этому универсальные модели GTM могут быть использованы для построения классификационных ландшафтов, осуществляющих разделение активных от неактивных соединений, для произвольных соединений и свойств, включая не участвовавших в их построении. В качестве примера на Рис. 32 приведены карты, построенные на базе универсальной модели GTM для активности химических соединений по отношению к рецепторам GPCR. Очевидно, отображение новых соединений на такие карты позволяет прогнозировать типы их биологической активности.

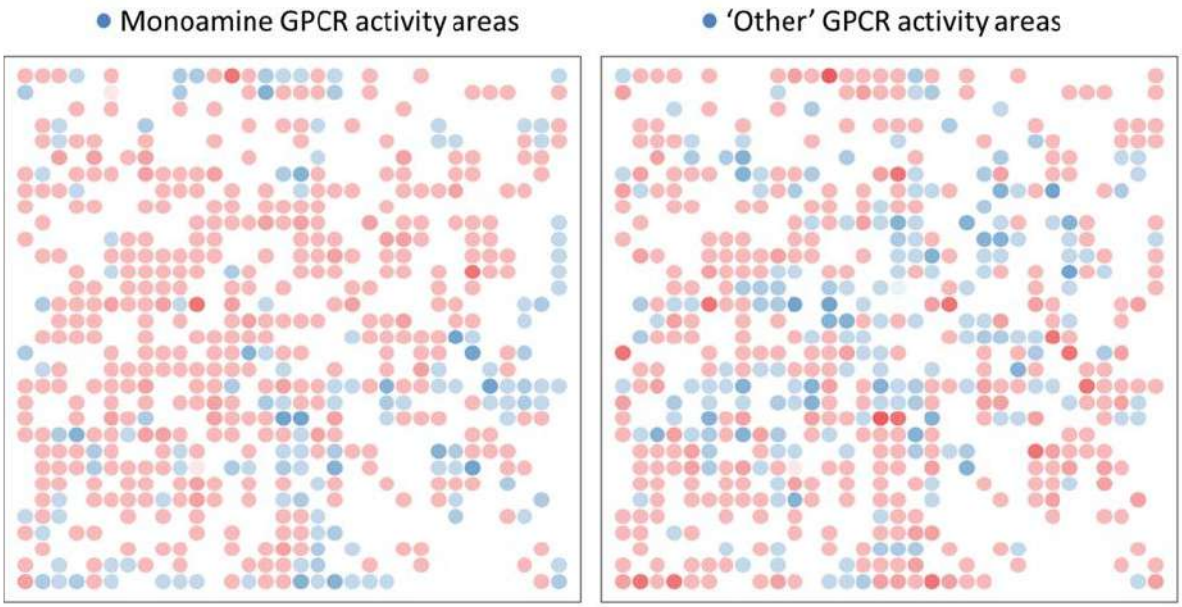


Рис. 32. Карты для лигандов рецепторов GPCR, построенные с помощью универсальных моделей GTM. Оттенки голубого обозначают связывание лигандов с рецептором, красного – отсутствие связывания, а интенсивность цвета – заселенность узла решетки. Рисунок из публикации [90] приводится с разрешения издательства. Copyright (2015) Springer.

2.2.3. Индексы SARI и SALI

Для того, чтобы охарактеризовать зависимость «структура-свойство» в химическом пространстве дескрипторов, было предложено использовать индексы SARI и SALI. Индекс SARI (*Structure-Activity Relationships Index*) [39] характеризует соотношение структура-активность (структура-свойство) глобально. Он вычисляется как сумма двух факторов: «оценки непрерывности» (англ. *continuity score*), $score_{cont}$, и «оценки разрывности» (англ. *discontinuity score*), $score_{disc}$:

$$SARI = \frac{1}{2}\left(score_{cont} + (1 - score_{disc})\right) \quad (30)$$

«Оценка непрерывности» $score_{cont}$ характеризует взвешенное по активности структурное разнообразие. Для его расчета сначала вычисляется для него «сырое» (первоначальное) значение, raw_{cont}, по формуле:

$$raw_{cont} = 1 - \frac{\sum_{i<j} w_{ij} sim(i,j)}{\sum_{i<j} w_{ij}}, \quad (31)$$

где $sim(i,j)$ – мера сходства молекул i и j, которую предлагается в оригинальной публикации оценивать как значения индекса Танимото для «молекулярных отпечатков» MACCS этих молекул. Значение веса w_{ij} вычисляется для всех пар молекул (i, j) по формуле:

$$w_{ij} = \frac{act_i \times act_j}{1 + |act_i - act_j|}, \quad (32)$$

где act_i и act_j – значения активности для молекул i и j, соответственно.

«Оценка разрывности» характеризует среднюю разницу значений активности у пар близких молекул. Для нее «сырое» значение вычисляется по формуле:

$$raw_{disc} = \frac{\sum_{\{i,j:sim(i,j)>0.6, i<j\}} \left|act_i - act_j\right| \times sim(i,j)}{|\{i,j: sim(i,j) > 0.6, i < j\}|} \quad (33)$$

где суммирование ведется по всем парам молекул i и j, мера сходства для которых превышает 0.6. Число таких пар стоит в знаменателе этой дроби.

Далее «сырые» значения оценок переводятся в z-оценки (англ. Z-scores) путем вычитания их среднего по выборке значения *mean* и деления на соответствующее стандартное отклонение *sd*:

$$zscore_{cont} = \frac{raw_{cont} - mean(raw_{conr})}{sd(raw_{conr})} \quad (34a)$$

$$zscore_{disc} = \frac{raw_{disc} - mean(raw_{disc})}{sd(raw_{disc})} \quad (35b)$$

После этого для приведения к единому интервалу значений применяется кумулятивная функция распределения для нормального распределения Ф:

$$score_{cont} = \Phi(zscore_{cont}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{zscore_{cont}} exp\left(-\frac{1}{2}t^2\right) dt \quad (36a)$$

$$score_{disc} = \Phi(zscore_{disc}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{zscore_{disc}} exp\left(-\frac{1}{2}t^2\right) dt \quad (37b)$$

Индекс SARI принимает значения от 0 до 1. Низкие значения индекса SARI свидетельствует о том, что зависимость структура-активность не является гладкой и содержит значительное число «разрывов», тогда как высокие значения SARI – о гладкости зависимости структура-активность, когда сходные по значениям дескрипторов химические структуры обладают близкой активностью.

Индекс SALI (*Structure-Activity Landscape Index*) [91] является локальным. Он вычисляется отдельно для пар близких молекул и используется для того, чтобы численно описать «*рифы активности*» (англ. *activity cliffs*) – участки в химическом пространстве дескрипторов, в которых небольшая вариация в их значениях приводит к большому изменению свойств химических соединений [92]:

$$SALI_{i,j} = \frac{|act_i - act_j|}{1 - sim(i,j)} \quad (38)$$

3. БИБЛИОТЕКИ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

Библиотека соединений, содержащая даже миллионы соединений, не способна покрыть все химическое пространство возможных соединений. По нашим оценкам существует около 10^{33} молекул, «похожих» на лекарства [93]. Некоторые другие опубликованные в литературе оценки, основанные на менее жестких критериях, достигают астрономического числа 10^{60} [94]! Миллионы соединений, созданных человеком, кажутся незаметной песчинкой на этом фоне. С другой стороны, ситуация облегчается тем фактом, что с точки зрения принципа сходства, сходные соединения с большей вероятностью проявят близкую активность, чем непохожие. Поэтому, если требуется улучшить свойства существующего соединения – надо искать лишь в его ближайшем окружении в химическом пространстве, если же требуется найти принципиально новые соединения – надо идти в более удаленные зоны химического пространства.

Дизайн библиотек соединений сводится к решению следующей задачи: как с помощью минимального числа испытаний найти вещество, обладающее желаемыми свойствами? Рациональный выбор соединений в библиотеку должен дать относительно небольшое число соединений, которые необходимо синтезировать и скринировать (или использовать в виртуальном скрининге). При этом зачастую также преследуется дополнительная цель: сделать так, чтобы вероятность того, что претенденты будут отброшены после скрининга вследствие проявления нежелательных эффектов, была минимальной.

3.1. ВИДЫ БИБЛИОТЕК СОЕДИНЕНИЙ

Гипотетически наиболее надежный способ найти соединение с требуемыми свойствами – перебрать все возможные варианты. Если количество исследуемых претендентов каким-то образом ограничено (например, только теми соединениями, которые можно получить в данной лаборатории) и оно невелико, то тотальная проверка всех возможных соединений в ходе виртуального или высокопроизводительного скрининга вполне осуществима. Однако, если число соединений очень велико, то такой подход «грубой силы» уже нереален. С учетом принципа сходства эта задача может быть разбита на два этапа – на первом этапе проводить поиск среди максимально различающихся соединений, найти среди них наиболее

многообещающие (называемые *хитами*, *кандидатами* или *лидерами*[1]), а на втором этапе проводить поиск только среди структур, максимально похожих на хиты, выявленные на первом этапе Рис. 33).

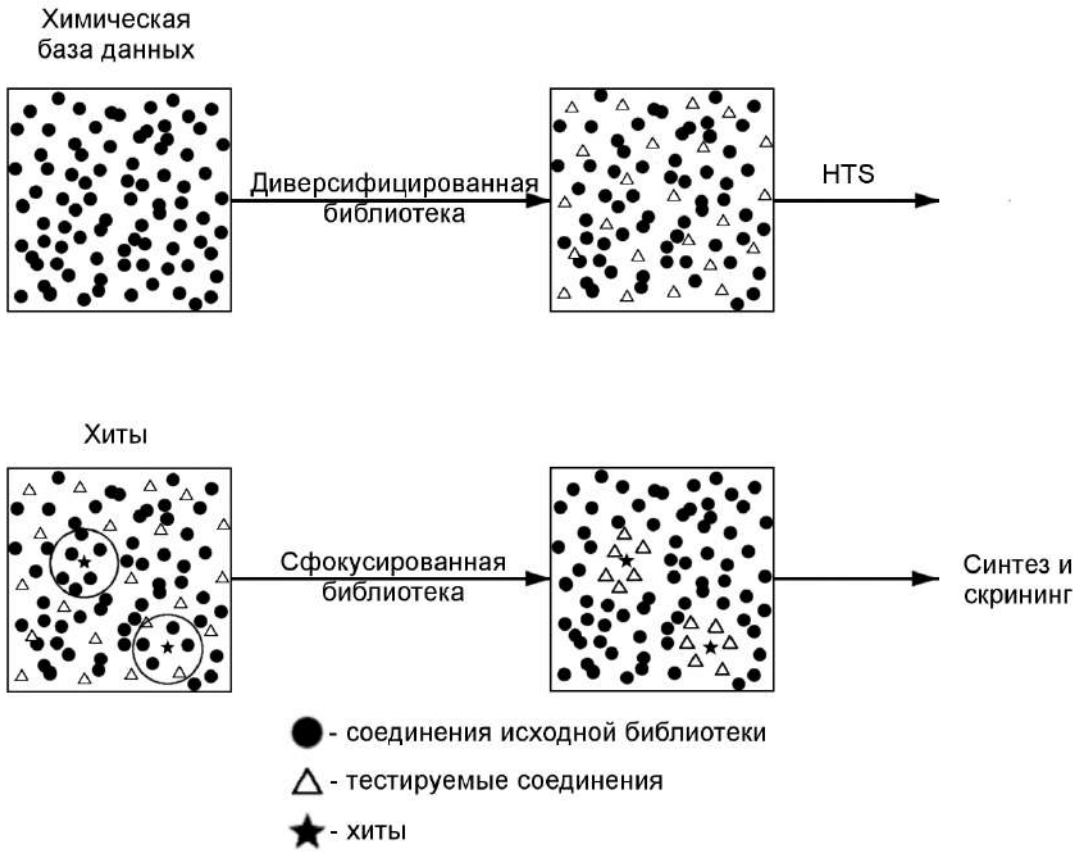


Рис. 33. Типичный алгоритм двухступенчатого поиска в большой базе соединений.

Таким образом, в зависимости от поставленной цели можно осуществлять две основные стратегии скрининга: диверсифицированный и сфокусированный скрининг. Соответственно, существуют *диверсифицированные* (или разбросанные, англ. *diverse*

[1] Между этими понятиями, строго говоря, есть некоторое отличие. Обычно под хитом понимают соединение, которое в данном биологическом тесте (bioassay), либо в виртуальном скрининге проявило желаемое свойство. Лидер – соединение, чья биологическая активность была подтверждена специальным, достаточно точным экспериментом (а не достаточно грубым, но быстрым тестом). Кандидат – соединение, которое в целом проявляет не только требуемые фармакодинамические, но и достаточно хорошие фармакокинетические свойства.

libraries) и *сфокусированные* (или целевые, англ. *focused libraries* или *targeted libraries*) библиотеки соединений.

Библиотеки могут быть созданы двумя способами: либо отбором из заданного набора соединений, либо путем генерации структур новых химических соединений, на основании, например, спецификации комбинаторной библиотеки, указания исходных реагентов для синтеза, требований к структуре или потенциальной синтетической доступности и др. Соответственно, в первом варианте пользователь имеет дело с соединениями, занесенными в какую-то базу данных, во втором случае такой информации нет, однако может иметься набор соединений, из которых посредством определенных реакций нужно получить соединения библиотеки. Соединения, занесенные в базу, могут быть реально известными и описанными, либо сгенерированными на компьютере.

В диверсифицированном варианте осуществляют отбор как можно более разнообразных и непохожих друг на друга соединений с целью охватить как можно большую область химического пространства и найти новые типы соединений с заданными свойствами. При сфокусированном биологическом скрининге, наоборот, используют библиотеки родственных или похожих соединений, что позволяет, например, оптимизировать активность препарата: зная приблизительную структуру активной молекулы, выбрать оптимальный вариант, обладающий наименьшим числом нежелательных эффектов.

Аналогичные подходы могут реализовываться в виртуальном скрининге: для реализации диверсифицированной стратегии скрининга используется предельно разнообразный набор существующих или сгенерированных на компьютере соединений, а для сфокусированной – библиотеки соединений, сгенерированных с использованием средств теоретической комбинаторной химии на основе заданного общего каркаса.

Для создания диверсифицированных библиотек необходимо из достаточно большой базы соединений отобрать те, которые максимально не похожи друг на друга. Это довольно серьезная проблема, часто стоящая перед фармацевтическими компаниями, которая может быть решена только средствами хемоинформатики. Предельно разнообразный набор дает больше шансов, что соединение с требуемым типом активности будет найдено с минимальными затратами. Отбор предельно разнообразных соединений из базы в миллионы соединений невозможно сделать вручную – для этого

применяют специальные методы отбора соединений в диверсифицированные библиотеки.

Создание сфокусированных библиотек обычно требует предварительного отбора некоторого числа базовых соединений, «вокруг» которых ведется генерация библиотеки. Для скрининга сфокусированная библиотека создается так, чтобы созданные соединения были похожи на уже известные препараты. Например, в методе Focus2D, разработанном в группе А. Тропши, соединения генерировались стохастически (метод симулированного отжига) путем комбинирования фрагментов. В процессе генерации проводилась приоретизация получаемых соединений с учетом евклидовых расстояний между ними в пространстве дескрипторов [95].

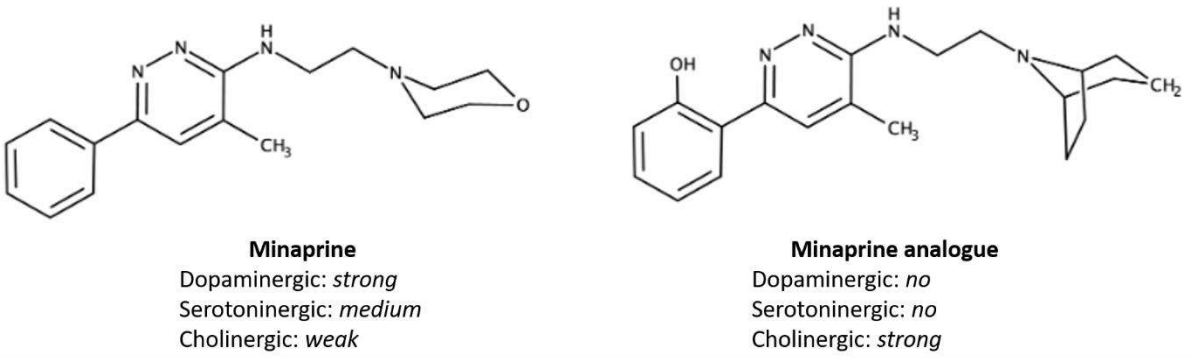


Рис. 34. Изменение структуры антидепрессанта минаприна приводят к тому, что его побочный эффект ингибирования становится доминирующим [98].

В качестве базового соединения, на который ориентируются при создании сфокусированных библиотек, не обязательно берется претендент, выявленный на предыдущем этапе скрининга, или хорошо зарекомендовавшее себя лекарственное средство. За основу может приниматься препарат, который обладает интересующими нас свойствами как побочными. Способ оптимизации, при котором модификация химической структуры приводит к тому, что побочная активность лекарственного препарата становится его основной терапевтической активностью, называется *селективной оптимизацией побочных эффектов* (SOSA, англ. selective optimization of side activity) [96, 97]. К примеру, антидепрессант минаприн (Рис. 33) обладает слабым эффектом ингибирования ацетилхолинэстеразы. Небольшая модификация его структуры приводит к кардинальной смене активности – полученное соединение является специфичным ингибитором холинэстеразы [98, 99], что позволяет использовать его

для лечения слабоумия и болезни Альцгеймера. При этом способность к ингибированию мускариновых рецепторов, обусловливавших активность минаприна, полностью теряется.

Как и в случае химического сходства, понятие химического разнообразия не имеет четкого определения и меры, поэтому было предложено множество подходов к выбору максимально разнообразного поднабора соединений. Так же, как и в анализе подобия, определение химического пространства для определения разнообразия соединений базируется на использовании молекулярных дескрипторов. В химическом пространстве дескрипторов мерой отличия двух соединений друг от друга является расстояние между ними. Следовательно, задача создания *диверсифицированной библиотеки соединений* сводится к нахождению их поднабора, входящие в который соединения максимально удалены в химическом пространстве друг от друга. Вместе с тем, строгий метод полного перебора всех возможных комбинаций наборов, состоящих из n соединений базы данных, содержащей N молекул, является крайне неэффективным. Число возможных вариантов таких наборов равно

$$\frac{N!}{n!\,(N-n)!}$$

Можно подсчитать по этой формуле, что существует более 10 миллиардов вариантов выбора наборов из 10 соединений в базе из 50 соединений. Обычно же в хемоинформатике приходится иметь дело с существенно большим числом соединений – вплоть до баз из миллионов соединений. Следовательно, для формирования диверсифицированных библиотек необходимо использовать нестрогие методы отбора из базы данных наборов разнообразных соединений. Наиболее популярные из них описаны ниже.

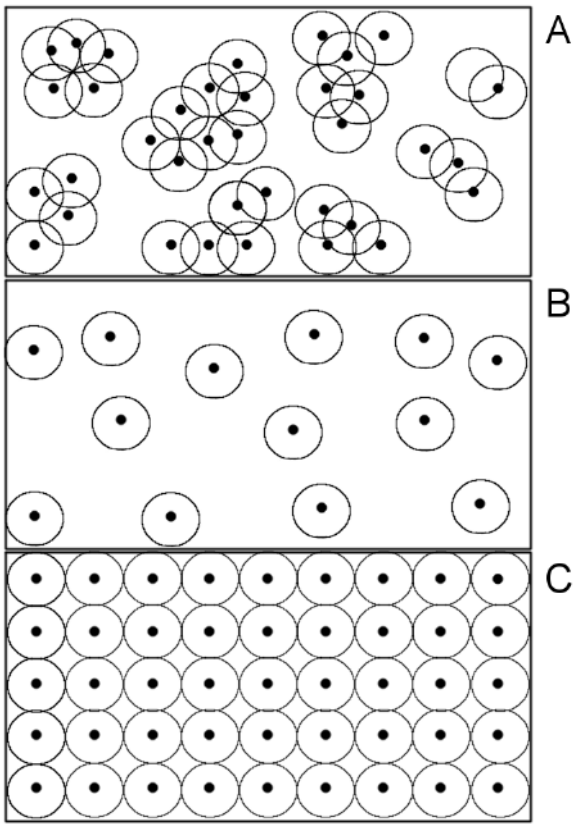


Рис. 35. Возможные проблемы при формировании диверсифицированных библиотек соединений в химическом пространстве. A - слишком кучно расположенные молекулы и наличие непокрытых областей, B - слишком редко расположенные молекулы и много непокрытых областей, C – идеальный случай. Точками обозначены соединения, кругами – области, внутри которых содержатся соединения похожие на них по свойствам.

Параметры, используемые при создании диверсифицированных библиотек, могут существенно влиять на их качество. При использовании сформированных библиотек в виртуальном скрининге максимальный шанс найти активное соединение с минимальными временными или финансовыми затратами, приходящимися на каждое соединение, получается, когда:

- отбираются разнообразные соединения, которые должны быть отдалены друг от друга (см. в качестве контрпримера Рис. 35A),
- отсутствуют общие ближайшие соседи для двух отобранных соединений, то есть отобранные соединения не «кучкуются» (см. в качестве контрпримера Рис. 35B),

- объем областей химического пространства, не представленных ни одним соединением, должен быть минимальным (см. в качестве примера Рис. 35С).

Вместе с тем необходимо достигать баланса между сфокусированностью и диверсификацией библиотеки, к какому виду бы она не относилась. В случае сфокусированных библиотек проблема соблюдения баланса диктуется тем, что существует (потенциально) огромное количество соединений, похожих на заданное, и необходимо даже в сфокусированных библиотеках создавать определенное разнообразие для того, чтобы наиболее эффективно использовать ограниченное число испытаний. В случае диверсифицированных библиотек необходимо принимать во внимание известную информацию о природе биологической цели, а также соображения о синтетической доступности соединений.

3.2. КОМПОНЕНТЫ ПРОЦЕДУРЫ ДИЗАЙНА БИБЛИОТЕК

Дизайн библиотеки соединений может включать три компонента (но не обязательно все): (1) генерация соединений; (2) отбрасывание соединений, не удовлетворяющих определенным правилам; (3) отбор соединений.

Создание библиотек (особенно для биологического скрининга) требует учета множества факторов. Эти факторы связаны с активностью соединения по отношению к данной биологической мишени или семейства мишеней, растворимостью, особенностями их абсорбции организмом (насколько легко вещество проникает в организм), распределением, метаболизмом и выведением, токсичностью, экономическими соображениями (доступностью реагентов, возможностью оптимизации синтеза, возможностью очистки соединения). Желательно не включать в библиотеку соединения, которые могут проявлять токсичность (в том числе тератогенность, мутагенность), будут плохо всасываться организмом, образовывать токсичные метаболиты, необратимо связывать биомишень, а также соединения, которые сложно синтезировать. Желание следовать этим требованиям привело к возникновению таких понятий как сходство с молекулами лекарств (употребляется также термин «лекарствоподобие», англ. *drug-likeness*) и сходство с соединением-лидером («лидероподобие», англ. *lead-likeness*). По этой причине дизайн экспериментальных библиотек часто включает фильтрование библиотеки от соединений, которые могут не пройти последующие (после виртуального скрининга) тесты и не

удовлетворяют определенным правилам (например, правилу Липинского, сходству с соединением-лидером, а также имеют химически активные группы). Эта процедура зачастую является одной из ранних стадий виртуального скрининга.

Отбор соединений является ключевой задачей при формировании их библиотек. Отбор соединений для сфокусированных библиотек обычно не представляет серьезной проблемы: похожие соединения либо создаются методами «экспериментальной» (т.е. реальным синтезом) или «теоретической» (т.е. комбинаторным перечислением на компьютере, см. раздел 3.5.1) комбинаторной химии на основе общего каркаса, либо отбираются из баз данных с использованием поиска по сходству. В то же время отбор соединений в диверсифицированную библиотеку представляет достаточно сложную задачу. Эта процедура, как правило, занимает основное время, требующееся на дизайн библиотек. Ее продолжительность быстро растет с увеличением числа соединений как в исходной базе данных, так и в формируемой библиотеке. Тем не менее, она крайне важна для удешевления проведения высокопроизводительного скрининга. Сформированные диверсифицированные библиотеки существующих соединений представляют по этой причине особую коммерческую ценность.

Начиная с 2013 года, осуществляется общеевропейский проект EU-Openscreen, направленный на создание диверсифицированного набора из 100 000 «лекарствоподобных» соединений для последующего широкомасштабного биологического скрининга и поиска новых лекарств. Для реализации этого проекта на первом этапе 5 различных исследовательских групп создали с использованием различных технологий библиотеки, содержащие по 40 000 наиболее разнообразных соединений, которые потом были объединены в один общий набор из 200 000 соединений. Из этого набора 100 000 соединений будут закуплены в скрининговую библиотеку, которая может стать коммерческим продуктом.

Вычислительная сложность создания диверсифицированных библиотек приводит к тому, что для проведения виртуального скрининга эту процедуру используют относительно редко, стараясь заменить фильтрами или быстрыми методами отбора, такими как случайный выбор соединений из исходной базы данных.

3.3. ГЕНЕРАЦИЯ НА КОМПЬЮТЕРЕ СОЕДИНЕНИЙ ДЛЯ СКРИНИНГОВЫХ БИБЛИОТЕК

Для проведения виртуального скрининга необходимо создание на компьютере библиотеки скринируемых соединений. Для виртуального скрининга могут использоваться как уже синтезированные соединения, собранные в базе данных, если нужно найти активное соединение среди существующих, так и сгенерированные теоретически, если целью является поиск еще не синтезированных соединений.

Проблема генерации структур химических соединений состоит в том, что их число практически бесконечно, и их генерация «вслепую» (без каких-либо заданных ограничений) является крайне неэффективной процедурой, поскольку огромное количество генерируемых при этом соединений имеет очень мало шансов пройти последующий отбор в виртуальном скрининге. Включение «химической информации» в процесс генерации соединений (виртуальный синтез) может привести к: (i) получению синтетически доступных структур, (ii) созданию библиотек с большой долей соединений с требуемыми характеристиками, (iii) уменьшению пространства поиска до приемлемого размера. «Химическая информация» включается в процесс генерации структур соединений, как правило, за счет использования определенных фрагментов, которые служат строительными блоками, а также за счет указания того, как эти блоки должны соединяться между собой.

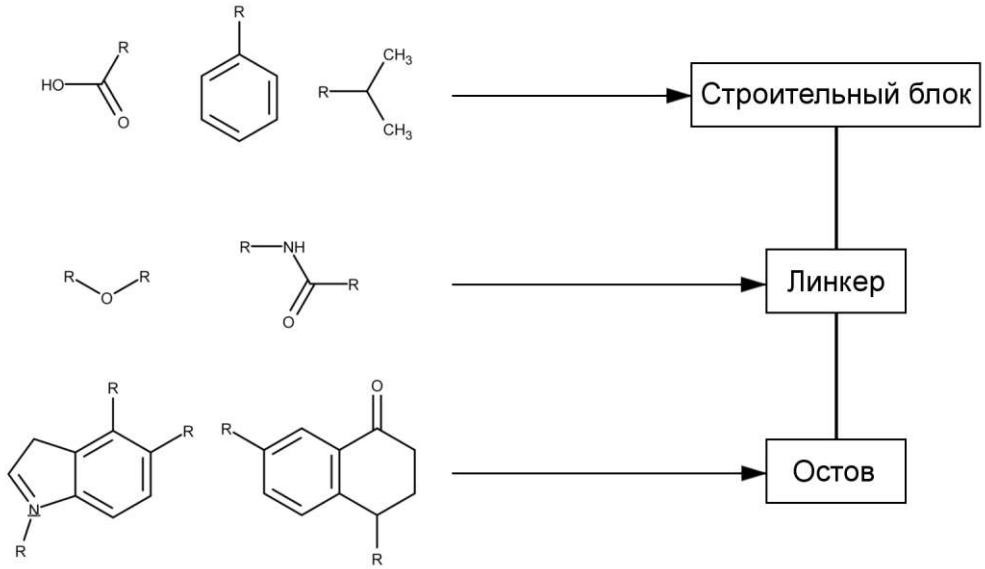


Рис. 36. Схема объединения различных типов фрагментов в комбинаторную библиотеку.

Обычно выделяют три типа фрагментов, из которых может проводиться сборка молекул: *остов* (скаффолд, каркас, темплат), *линкер* и *строительный блок* (Рис. 36). *Остов* – это фрагмент, являющийся общим для всех соединений библиотеки. У остова может быть одна или несколько точек присоединения фрагментов, причем их положение на остове может быть как фиксировано, так и варьируемо. Выбор остова может определяться его синтетической доступностью (наличием удобного способа его синтеза) и активностью соединений, содержащих этот фрагмент (например, стероидный, пенициллиновый фрагмент). *Линкер* – это группа, которая соединяет два или большее количество фрагментов в одну структуру. Линкер имеет два или, реже, большее число точек присоединения фрагментов. Как правило, линкер не ответственен за связывание с белком и служит для того, чтобы зафиксировать требуемые фрагменты в определенном положении или разнести их на определенное расстояние, увеличить или уменьшить конформационную гибкость. *Строительный блок* – это варьируемый фрагмент, присоединяемый к каркасу. Как правило, он содержит одну точку присоединения фрагментов. Они могут быть выбраны из каких-либо теоретических соображений относительно активности, способности к связыванию с теми или иными группами в ферментах, синтетической доступности. Генерация структур может осуществляться как без указания каркаса или линкера, так и с использованием нескольких каркасов, линкеров или строительных блоков.

3.3.1. Генерация библиотеки соединений

Большинство способов генерации структур химических соединений для включения в библиотеку для скрининга относятся к четырем основным категориям: полное перечисление, детерминистский и недетерминистский подходы, биоизостерное замещение.

При полном перечислении молекулярные графы получаются соединением фрагментов между собой во всех возможных комбинациях. Этот способ, однако, может привести к комбинаторному взрыву – даже из небольшого числа фрагментов могут быть получены миллионы (если не миллиарды) молекул. С другой стороны, это позволяет очень быстро получать огромные виртуальные библиотеки соединений. Для генерации соединений на первом этапе необходимо определить библиотеку фрагментов, из которых будет происходить сборка соединений. Критический отбор фрагментов является ключевой

стадией, поскольку позволяет избежать генерации непрактично большого числа структур, формировать библиотеки синтетически доступных и перспективных с точки зрения скрининга соединений. Полное перечисление соединений на втором этапе осуществляется с помощью методов маркирования фрагментов, реакционных трансформаций или с использованием структур Маркуша.

Чтобы избежать комбинаторного взрыва, используют подходы, в основе которых лежит оптимизация определенной целевой характеристики, которая в какой-то мере должна отражать «качество» сгенерированных структур. Преимущество этого состоит в том, что не тратятся ресурсы на генерацию и последующий скрининг заведомо неподходящих структур. В качестве такой характеристики может выступать, например, энергия взаимодействия с сайтом связывания макромолекулы белка, играющего роль биологической мишени, сходство с молекулой известного лекарства и другие важные характеристики. Сама оптимизация при этом может вестись в рамках детерминисткого и недетерминистского подходов.

В детерминистских подходах процесс объединения фрагментов в молекулу однозначно определяется тем, насколько молекула и ее фрагменты удовлетворяют определенным характеристикам. Детерминистский подход формально можно рассматривать как сочетание виртуального скрининга и дизайна библиотек. Рассмотрим в качестве примера метод FlexNovo [100, 101], в котором молекулы конструируются в полости рецептора из отдельных фрагментов. На первом этапе фрагменты размещаются в полости белка, так чтобы иметь максимально выгодные взаимодействия с аминокислотными остатками, т. е. осуществляется *докинг* (англ. docking) фрагментов в полость белка. Далее, начиная с определенного числа наиболее хорошо сдокированных (т. е. обладающими наилучшими значениями специальной скоринг-функции – некоторого подобия энергии взаимодействия) фрагментов, они последовательно объединяются в молекулы. На каждом шагу такого объединения определяется, насколько хорошо докируется объединенный фрагмент. Таким образом, каждый шаг объединения фрагментов в молекулы определяется некоторой характеристикой получаемого на этом шаге объединенного фрагмента.

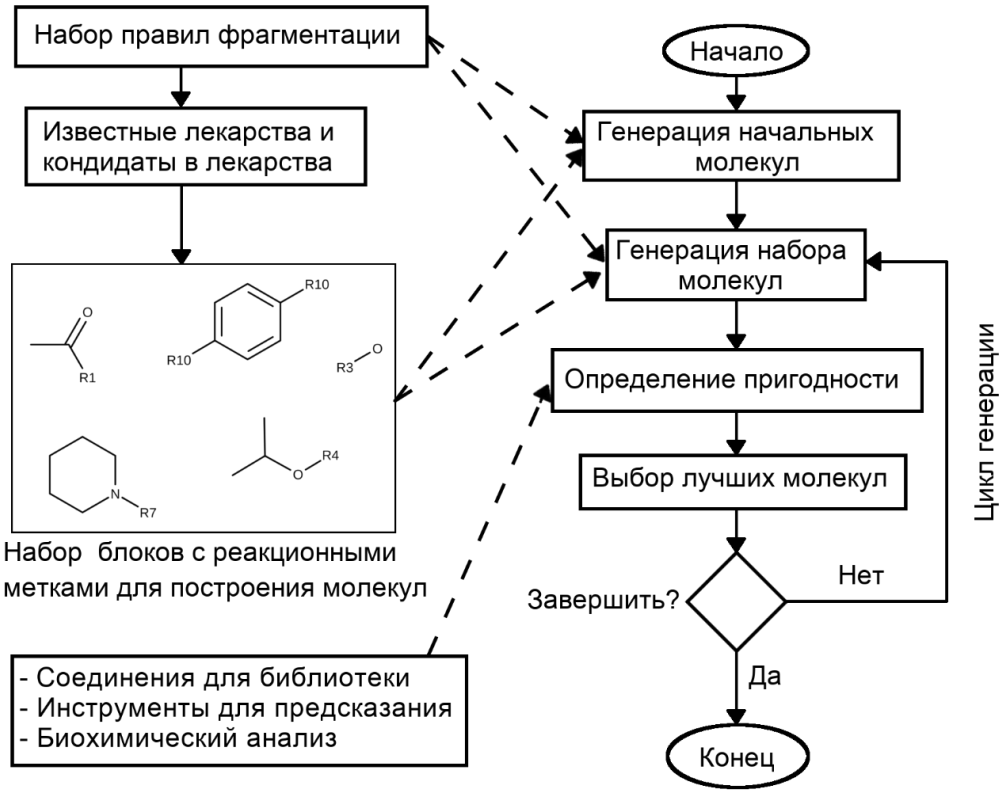


Рис. 37. Алгоритм стохастической генерации молекул с использованием генетического алгоритма в программе FLUX.

В недетерминистских подходах генерация химических структур осуществляется с помощью стохастических методов, в которых характеристика «качества» структуры определяет вероятность, с которой она может быть сгенерирована. Числовые значения такой характеристики определяют особенности наборов генерируемых структур. Например, задание характеристики, связанной с оценкой сходства молекул, определяет, будет ли сгенерирован диверсифицированный либо сфокусированный набор соединений. В качестве примера недетерминированного подхода можно привести метод FLUX [102, 103], в котором генерируется набор молекул, сходных с заданной (называемой шаблоном, темплатом, англ. *template*). Генерация ведется с помощью генетического алгоритма оптимизации, в котором характеристика «качества» используется в качестве *функции соответствия* (англ. fitness function), Рис. 37. На первом этапе работы генетического алгоритма генерируется путем комбинирования фрагментов набор из определенного числа соединений («особей», «хромосом»). Далее эти соединения ранжируются по значению функции соответствия, в качестве которой в методе FLUX используется мера сходства (индекс Танимото или

евклидово расстояние) с шаблоном. Некоторое количество молекул, наиболее похожих на шаблон (в терминах генетического алгоритма – наиболее «жизнеспособных особей»), становится «родителями» следующего поколения молекул, тогда как остальные «умирают» (то есть не рассматриваются далее). Важно подчеркнуть, что отбор при этом осуществляется стохастически (т. е. случайным образом) в соответствии с вероятностями, определяемыми значениями функции соответствия (в данном случае сходством с шаблоном). «Родители» дают новое поколение молекул с помощью осуществляемых также случайным образом процедур *кроссинговера* (англ. cross-over), когда две молекулы обмениваются какими-то фрагментами, и мутаций, когда один фрагмент заменяется на любой другой из библиотеки фрагментов. «Родители» после этого «умирают» (в программе FLUX не используется режим «элитарности», когда родители с наилучшими значениями функции соответствия «выживают»), а для «потомков» вычисляется функция соответствия, и процедура отбора вновь повторяется уже на следующем этапе эволюции. Итерации продолжаются до тех пор, пока не будет достигнут заданный критерий остановки (в программе FLUX таким критерием служит количество итераций (число поколений в эволюции). На недетерминистском подходе основан также метод SYNOPSIS, однако в нем генерация молекул осуществляется не путем соединения фрагментов между собой, а путем применения 70 типов хорошо известных синтетических трансформаций к соединениям из базы данных [104]. Оптимизация структур молекул в этом случае ведется с использованием функции соответствия и стохастических методов «искусственного отжига» и генетического алгоритма. В методе Molecule Evaluator [105] молекулы также собираются с помощью генетического алгоритма, но не из фрагментов (как в программе FLUX), а из отдельных атомов с помощью стохастических операций изменения (мутаций) структур молекул (добавлением, перемещением атомов, изменением типов атомов и связей и др). В методе LEA3D с использованием генетического алгоритма осуществляется генерация трехмерных структур молекул из 3D фрагментов с использованием комбинированной функции соответствия, включающей скоринг-функцию для докинга лиганд-белок [106].

Методы биоизостерного замещения исходят из предположения, что фрагменты молекул, характеризующиеся одинаковой пространственной формой и одинаковым распределением электронных характеристик, будут одинаковым образом связываться с биологической мишенью. Если это так, то генерировать новые

соединения, похожие на заданные (то есть создавать сфокусированные библиотеки соединений), можно, замещая такие фрагменты друг на друга. *Биоизостеры – это заместители или группы с близкими физическими и химическими свойствами, которые используются для модификации химических соединений без существенного изменения биологических свойств* (Рис. 38). Биоизостерное замещение используется для модификации соединений с целью исключить какие-то нежелательные свойства соединения (токсичность, низкая селективность действия, метаболическую нестабильность и др.), заменить дорогие или сложно синтезируемые фрагменты на более доступные, уйти от ранее запатентованной структуры.

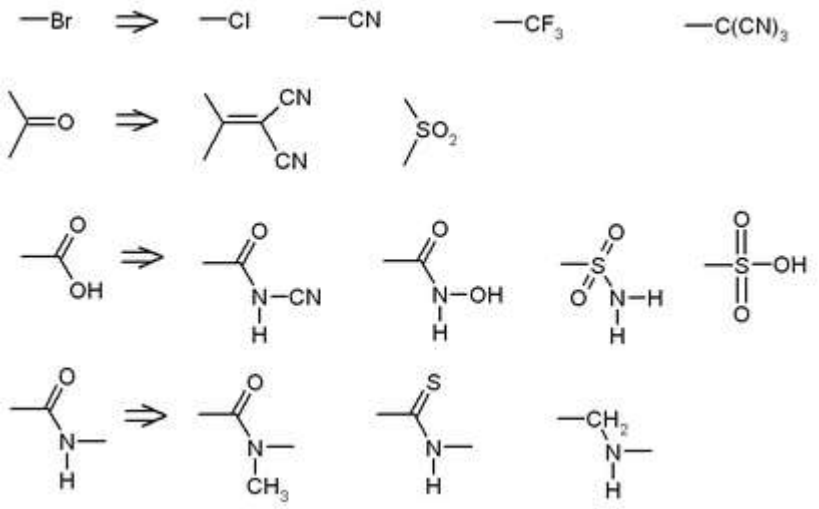


Рис. 38. Примеры биоизостерных функциональных групп

Программный пакет ReCore [107] от BioSolveIT производит биоизостерное замещение указанных пользователем фрагментов молекулы с учетом геометрических ограничений, а также взаимодействий данного фрагмента с рецептором. Библиотека фрагментов для такого замещения может быть сформирована исходя из базы трехмерных структур соединений. Программа BROOD [108] от OpenEye производит замещение выбранных фрагментов в молекуле на основании их формы и электростатического поля, создаваемого ими. При этом может использоваться встроенная или специально созданная библиотека фрагментов. Подход, основанный на близости молекулярных полей, используется программой SparkV10 [109] от Cresset Group для генерации новых соединений. В этом случае биоизостерное замещение фрагментов (в том числе замещения атомов) производятся так, чтоб влияние фрагмента на создаваемое молекулой поле (рассматриваются электростатическое поле, поле гидрофобности и ван-дер-ваальсовых взаимодействий) было минимальным.

3.3.2. Формирование наборов фрагментов для генерации химических соединений

Ключевой стадией для сборки молекул из фрагментов для создания комбинаторной библиотеки является формирование наборов таких фрагментов. Совершенно очевидно, что фрагменты должны быть выбраны так, чтобы их можно было соединять между собой с помощью протекающих с высоким выходом химических реакций. Второе условие заключается в том, чтобы используемые в таких реакциях реагенты были доступными и по возможности дешевыми (что означает, что должны быть хорошо разработаны способы получения фрагментов). Фрагменты должны позволять собирать из них молекулы, обладающие определенным типом активности: они должны обладать определенными фармакофорными группировками, быть не слишком крупными (чтоб избежать нежелательных взаимодействий), не слишком гидрофобными. Для идентификации таких фрагментов сформулировано «правило трех» [110]: фрагменты для создания библиотек «лекарствоподобных» соединений должны обладать следующими характеристиками:

1. молекулярная масса фрагмента меньше 300,
2. число H-доноров и H-акцепторов не более 3,
3. гидрофобность (вычисленная по методу ClogP) не больше 3;

предлагается также использовать два дополнительных правила:

4. число вращающихся связей во фрагменте не более 3,
5. площадь полярной поверхности фрагмента не более 60 Å^2.

Традиционным способом формирования наборов фрагментов (строительных блоков и линкеров) является их выбор «вручную», исходя из синтетических возможностей лаборатории (наличия тех или иных синтонов[1] и доступных реакций) и знаний (гипотез) относительно структуры желаемой молекулы. Например, если известно, какова должна быть длина линкера, то могут быть перебраны все возможные структуры линкера, которые относительно легко могут быть получены в данной синтетической лаборатории и обладают заданной длиной. Этот подход, однако, с трудом может быть применен к созданию больших баз разнообразных соединений.

Другим подходом к формированию наборов фрагментов является «нарезка» соединений из баз, содержащих структуры синтезированных соединений. Для того, чтобы получаемые при рекомбинации

[1] Синтон – ключевой промежуточный реагент, получение которого позволяет синтезировать множество различных соединений.

фрагментов соединения можно было легко синтезировать, необходимо «резать» по тем связям, которые легче всего синтетически создать. Поскольку процесс «нарезания» по таким связям похож на ретросинтетический анализ (см. раздел 6.1.1 в пособии 5), то и автоматическое формирования набора фрагментов может быть осуществлено с помощью *псевдо-ретросинтетических правил*. При применении таких правил для извлечения фрагментов из баз данных, содержащих структуры биологически активных химических соединений:

- структуры получаемых фрагментов будут связаны с конкретным типом активности. В этом случае можно ожидать, что молекулы, собранные на основе таких строительных блоков, будут обладать активностью, характерной для «ключевого» фрагмента,
- обеспечивается синтетическая доступность генерируемых из таких фрагментов соединений
- возможно создание сфокусированных библиотек путем комбинирования фрагментов, извлеченных из соединений, обладающих нужным типом активности.

Наиболее широко распространенным методом псевдо-ретросинтетического разбиения молекул на фрагменты являются RECAP (*REtrosyntetic Combinatorial Analysis Procedure*) [111]. Также компанией ChemAxon был разработан метод Fragmenter, представляющий собой обобщение правил RECAP.

3.3.2.1. Метод RECAP

Метод RECAP позволяет формировать наборы фрагментов путем извлечения их из баз данных, содержащих структуры биологически активных соединений. Он использует 11 типов связей, по которым могут «разрезаться» молекулы на фрагменты (приведен также номер): 1 – амидная, 2 – сложноэфирная, 3 – аминная, 4 – связи в мочевинном фрагменте, 5 – эфирная, 6 – двойная связь, 7 – связь с азотом в четвертичном амине, 8 – связь между ароматическим азотом и алифатическим углеродом, 9 – лактамная, 10 – одинарная связь между двумя ароматическими кольцами, 11 – сульфонамидная (эти связи и некоторые другие приведены на Рис. 18). Этот набор типов связей был получен путем анализа наиболее распространенных реакций в синтетической органической химии. Например, связи типа 3 могут быть легко образованы нуклеофильным замещением или восстановительным аминированием. Если после разделения

получается маленький фрагмент (водород, метил, этил, пропил, бутил), то фрагмент не отсекается. Связи, находящиеся в кольце, также не затрагиваются. В полученных фрагментах тип разорванной связи помечается. Например, если разрывается четвертичная аминная связь (тип 7 разрываемой связи, Рис. 18), то атомы азота и углерода помечаются цифрой 7 (например, как изотопы с массой 1 в SDF файле), чтобы потом можно было при сборке собирать молекулы, объединяя одинаковые типы связей.

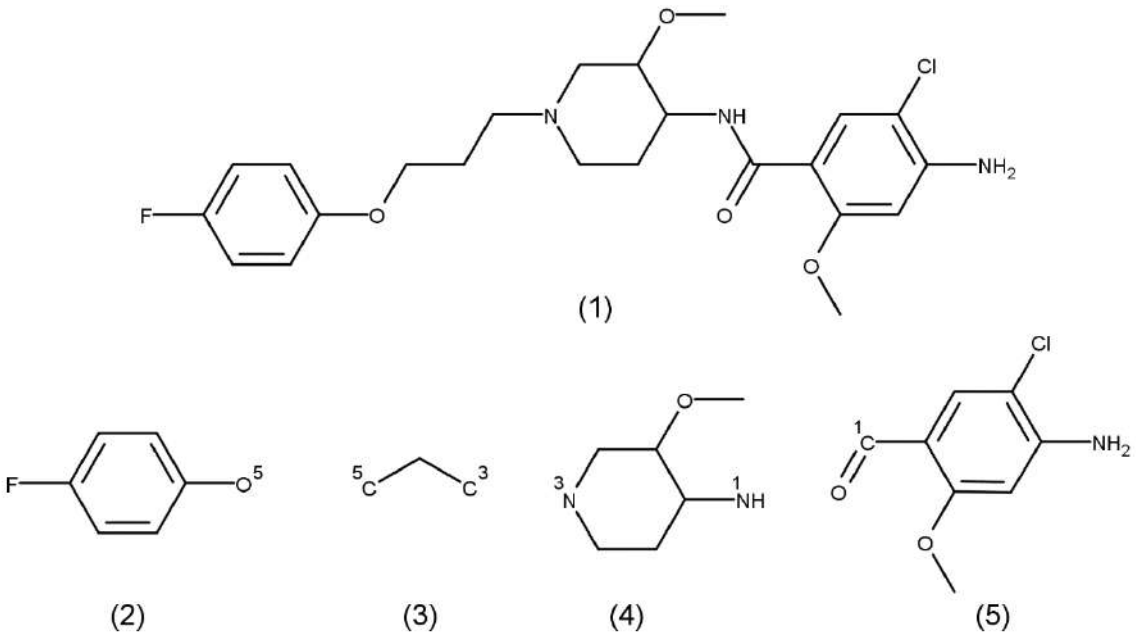


Рис. 39. Разделение молекулы гастрокинетика цизаприда с помощью RECAP.

Пример использования RECAP для разделения молекулы на фрагменты приведен на Рис. 39. Полученные с помощью RECAP фрагменты в большинстве случаев удовлетворяют «правилу трех» и широко используются для создания библиотек соединений. Этот метод, например, используется для создания и нарезки соединений в рассмотренном выше в разделе 3.3.1 методе FLUX. Метод RECAP имплементирован в библиотеке RDKit[1].

3.3.2.2. Метод Fragmenter (ChemAxon)

Fragmenter – это программа для осуществления разделения молекул на фрагменты по методу RECAP, разработанная в компании ChemAxon. В настоящее время она более не поддерживается, но

[1] Смотрите подробнее: https://www.rdkit.org/docs/source/rdkit.Chem.Recap.html

доступна в архивных версиях. Типы разрезаемых связей в программе Fragmenter те же, что и в RECAP, однако предусмотрена еще возможность добавлять собственные типы связей, задавая с помощью линейной нотации SMIRKS «реакции» деления молекулы на фрагменты. Во Fragmenter встроено 6 правил, частично присутствующих в оригинальном методе RECAP, позволяющих осуществлять настройку глубины фрагментации:

1. не разрезать связи в цикле (обязательное);
2. не разрывать связи с атомом водорода;
3. не разрывать связь атома углерода, находящегося в кольце, с гетероатомом (опциональное, можно отключать);
4. не разрывать связь, если один из получающихся фрагментов обозначен в специальном СТОП-списке;
5. не разрывать связь, если количество свободных валентностей у хотя бы одного полученного фрагмента превышает указанный лимит;
6. не разрывать связь, если число указанных явно атомов в любом из полученных фрагментов меньше указанной величины.

Если правила 1 и 2 соблюдаются в любом случае, то правила 3-6 можно отключить, и тогда для получения фрагментов будут разрезаться все ациклические связи, которые удовлетворяют любому из RECAP-типов разрываемых связей. Указывая определенные параметры, можно включать или выключать правила 3-6 в любой комбинации. Например, отключить правило 5 и 6 можно указанием нереалистичного лимита (скажем, 99). Это обеспечивает большую гибкость подхода по сравнению с оригинальным методом RECAP (Рис. 40). Программа также позволяет анализировать информацию о биологической активности соединений, в которых присутствуют получаемые фрагменты. Программа не поддерживается с января 2015 года, но доступна в архивных версиях модуля JChem.

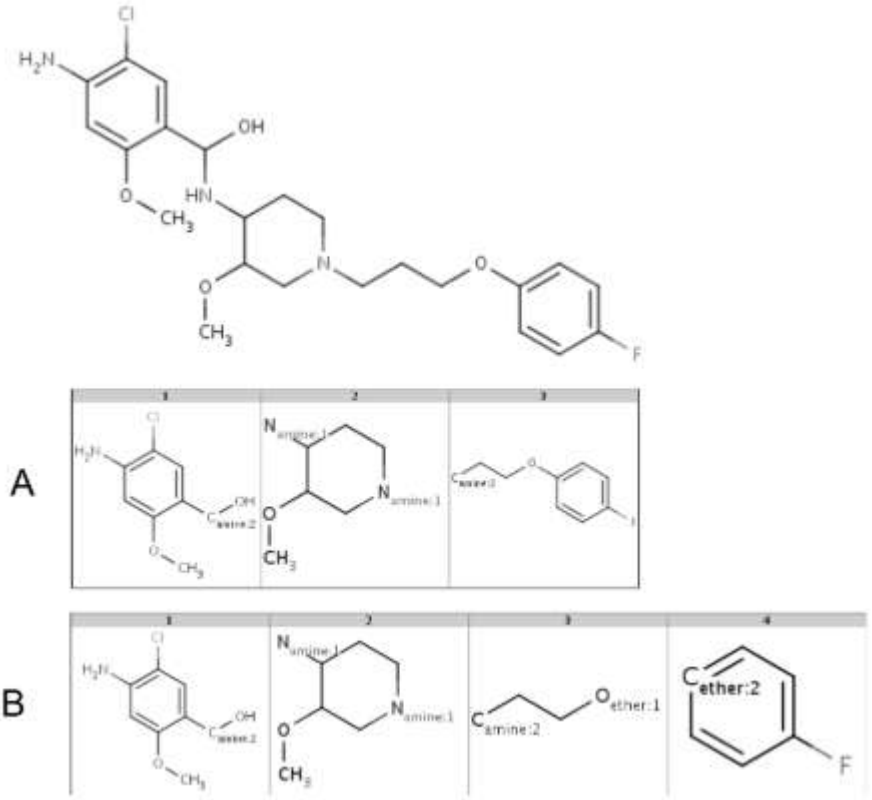


Рис. 40. Фрагменты, которые получаются из молекулы цизаприда с помощью программы Fragmenter, A – правило 3 включено, B – правило 3 отключено.

3.4. ОТБОР НАБОРА СОЕДИНЕНИЙ С ЗАДАННЫМ РАЗНООБРАЗИЕМ

Когда библиотека виртуальных соединений, которая может содержать миллионы и даже миллиарды молекул, создана, ключевой задачей для проведения последующего скрининга является выбор соединений. В этом случае необходимо отобрать соединения в библиотеку соединений определенного (как правило ограниченного) объема. Отобранная библиотека, какой бы она ни была, сфокусированной или диверсифицированной, должна обеспечивать баланс сходства и разнообразия молекул.

Отбор соединений в фокусированные библиотеки соединений не представляет особого труда, поскольку эффективные методы поиска по сходству хорошо разработаны. Кроме того, для отбора соединений в сфокусированные библиотеки используются методы, основанные на виртуальном скрининге.

Основной вычислительной сложностью создания диверсифицированных наборов является та, что проблема поиска

максимально различающихся соединений относится к NP-классу и, следовательно, требует применения специальных подходов. Важность решения проблемы поиска диверсифицированного набора соединений привело к разработке множества подходов к ее решению: методов на основе кластерного анализа; методов, основанных на различии; «клеточных» методов; методов оптимизации; планирования эксперимента. Далее мы остановимся на наиболее распространенных способах отбора максимально разнообразных наборов соединений.

Для создания библиотек максимально разнообразных соединений важно определить характеристики, позволяющие оценивать, насколько диверсифицированным получается набор. Задача отбора максимально разнообразных молекул в этом случае сводится к максимизации этой меры. В каждом методе используются свои меры разнообразия. Например, в кластеризации и методах отбора по несходству – это различные расстояния между кластерами и соединениями (см. далее). Уолдмэн с соавторами [112] на основе теоретических соображений о свойствах диверсифицированного набора предложил 5 признаков, которыми должна обладать функция, измеряющая разнообразие соединений в наборе:

1. Добавление в отобранный набор молекулы, расположенной в той же точке химического пространства, что и одна из присутствующих в нем молекул, не должно влиять на значение функции;
2. Добавление в отобранный набор любой новой молекулы (не совпадающей с уже присутствующими) должно приводить к увеличению функции;
3. Функция должна обладать пространство-заполняющим поведением, то есть она должна сильнее изменяться (увеличиваться) при заполнении пустых областей химического пространства, чем при заполнении густо заселенных областей;
4. Если дескрипторное пространство не бесконечно, то его заполнение бесконечным числом молекул должно стремиться к конечному значению функции разнообразия;
5. Если молекулы удаляются друг от друга, то это должно приводить к увеличению функции, однако если расстояние стремится к бесконечности, то функция при этом должна стремиться к некоторому конечному значению.

Сам М. Уолдмэн предложил использовать для этого функцию, основанную на вычислении площади под пересекающимися

гауссовыми функциями[1], размещенными в вершинах минимального охватывающего дерева (англ. *minimum spanning tree*) данного набора молекул. Минимальное охватывающее дерево, построенное на наборе молекул – это такой граф-дерево, вершинами которого являются выбранные молекулы, и при этом сумма длин ребер которого является минимально возможной. Тогда функция разнообразия равна площади под пересекающимися гауссовыми кривыми. Последняя вычисляется как сумма функций ошибок[2] на расстояниях между парами связанных вершин, Рис. 41.

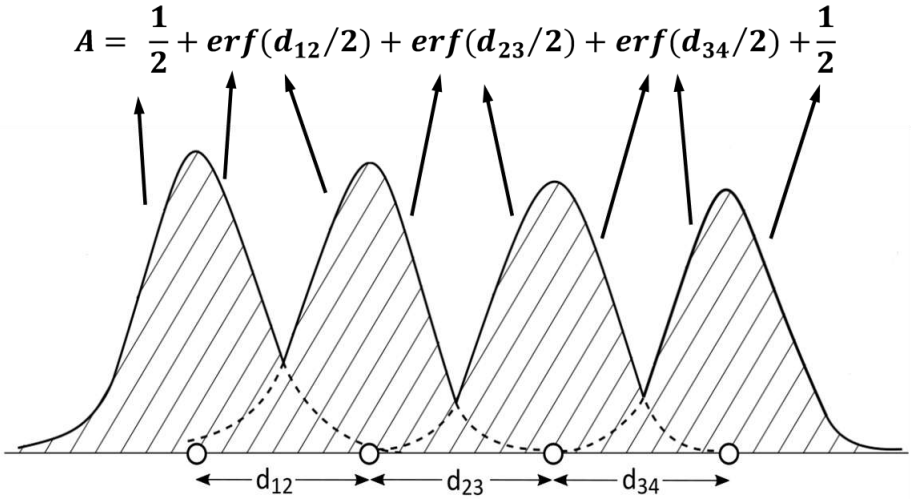


Рис. 41. Площадь под пересекающимися гауссовыми кривыми охватывающего дерева равна сумме функций ошибок. Рисунок адаптирован с работы [112].

---

[1] Гауссова функция - это математическая функция, задаваемая для одномерного случая формулой: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, где параметры μ и σ - вещественные числа. График функции имеет колоколообразную форму, положение максимума функции определяется параметром μ, а ширина – параметром σ. Функция Гаусса описывает важнейшее в статистике нормальное распределение вероятностей, с математическим ожиданием распределения μ и дисперсией σ.

[2] Функция ошибок (функция Лапласа, интеграл вероятностей) – это неэлементарная функция, возникающая в различных разделах математики, определяется как $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. С точностью до масштаба и сдвига совпадает с определенным интегралом функции Гаусса на интервале от нуля до заданного числа.

Большинство используемых для отбора молекул мер разнообразия не удовлетворяет тем или иным критериям, однако они, в отличие от функции разнообразия, предложенной Уолдмэном, могут быть легко вычислены. Поэтому функция Уолдмэна может быть использована для сравнения результатов различных методов отбора молекул. В то же время, для проведения эффективного отбора соединений она не годится и вместо нее применяют легче вычисляемые характеристики.

3.4.1. Отбор соединений с помощью методов кластерного анализа

Разнообразные методы кластерного анализа рассмотрены в разделе 2.14.1 пособия 4. Они позволяют сгруппировать соединения в кластеры (группы) таким образом, что соединения, принадлежащие одному кластеру, в большей степени схожи друг с другом, чем принадлежащие разным кластерам. Это позволяет сформировать представительный набор соединений, отбирая по одному (или небольшому количеству) из каждого кластера. Процесс отбора соединений с помощью кластерного анализа включает следующие шаги:

1. выбор дескрипторов для описания положения соединений в химическом пространстве;
2. вычисление расстояний между всеми соединениями в наборе;
3. использование алгоритма кластеризации для объединения молекулы в группы;
4. отбор искомых соединений в библиотеку выбором одного или нескольких молекул из каждого кластера.

3.4.2. Методы отбора, основанные на мере различия

В отличие от вышеупомянутых методов отбора, основанных на кластерном анализе, которые осуществляют отбор в два этапа, методы отбора, основанные на несходстве (различии) (DBCS, англ. dissimilarity-based compound selection), осуществляют отбор разнообразных соединений непосредственно, без формирования каких-либо промежуточных структур данных [113, 114]. Общий алгоритм методов этого типа, предложенный в 1969 году Р. Кеннардом и Л. Стоуном [115], можно описать следующим образом:

1. Выбрать соединение из библиотеки и поместить его в формируемый набор;
2. Пока количество соединений в наборе меньше требуемого:

2.1. Вычислить меру различия (например, расстояние в химическом пространстве дескрипторов) между соединениями из формируемого набора и оставшимися соединениями в библиотеке;

2.2. Выбрать из библиотеки следующее соединение такое, что оно будет максимально отличаться от отобранных, и поместить его в формируемый набор.

Существует большое разнообразие вариантов реализации этого алгоритма, отличающихся тем, каким образом выбираются соединения на первом шаге, каким образом вычисляется мера несходства молекул, а также на основании чего выбираются несхожие молекулы. Очевидно, что если на первом этапе отбирать соединение различным образом, то даже при прочих равных условиях полученные наборы будут отличаться. Таким образом, DBCS дает субоптимальное решение проблемы отбора – и успешное завершение процедуры совершенно не означает, что найденный набор молекул является наиболее разнообразным из всех возможных вариантов.

Существует несколько основных способов выявления первого соединения:

1. Выбирать соединение случайным образом,
2. Выбирать соединение, которое является наиболее представительным, то есть наиболее похожим на соединения библиотеки (имеющее максимальную сумму мер сходства с соединениями библиотеки),
3. Выбирать соединение, максимальным образом отличающееся от соединений библиотеки (имеющее максимальную сумму мер несходства с соединениями из библиотеки).

Для реализации шага 2 требуется определить меру несходства, на основании которой будут отбираться соединения в набор.

Поскольку на каждом шагу проводится расчет матрицы расстояний от n отобранных молекул до N молекул в библиотеке, то вычислительная сложность методов DBCS пропорциональна $O(n^2N)$. Учитывая, что обычно n выражается как доля от N (например, 1%), то время расчета растет пропорционально N^3, что не позволяет их использовать для больших баз. Тем не менее, существуют подходы, позволяющие существенно ускорить расчеты и отбирать наборы максимально несходных соединений из больших баз.

3.4.2.1. Алгоритмы максимального несходства

Одним из вариантов методов DBCS являются методы максимального несходства.

Поскольку задачей является отбор максимально несходных соединений, то в формируемый набор следует выбирать соединения, в наибольшей степени отличающиеся (т. е. максимально удаленные) от уже отобранных. Можно предложить несколько различных характеристик, описывающих, насколько данное соединение находится далеко от набора уже отобранных соединений (как в методах SAHN): минимальное (метод MaxMin) или максимальное расстояние (MaxMax) между данным соединением и уже отобранными молекулами, сумма этих расстояний (MaxSum) или их медиана (MaxMed). Не все они, однако, дают хорошие результаты: за исключением метода MaxMin все остальные виды расстояний приводят к появлению похожих соединений в формируемом наборе [116]. Поскольку это крайне нежелательно, то наиболее широко используются методы MaxSum и MaxMin [117]. Если уже отобрано m соединений, то в качестве меры удаленности i-ого соединения из библиотеки до отобранных соединений D_i берутся:

$$MaxSum: \quad D_i = \sum_{j=1}^{m} d_{ij} \qquad (39a)$$

$$MaxMin: \quad D_i = \min\left(d_{ij}\right), j = 1 \ldots m, \qquad (39a)$$

где d_{ij} – расстояние от i-ого соединения библиотеки до j-ого соединения из отобранного набора. Таким образом, метод MaxSum в качестве меры удаленности соединения i от отобранного набора использует сумму расстояний от него до отобранных соединений, а метод MaxMin – расстояние до ближайшего отобранного соединения.

Несмотря на то, что метод MaxSum обеспечивает отбор максимально удаленных друг от друга соединений (что может быть оценено по сумме расстояний между отобранными соединениями), среди отобранных соединений присутствует близко друг к другу расположенные (Рис. 42) [117]. Отобранные им соединения имеют тенденцию располагаются «по углам» химического пространства. Этого недостатка лишен метод MaxMin, обеспечивающий равномерный отбор соединений из разных областях химического пространства.

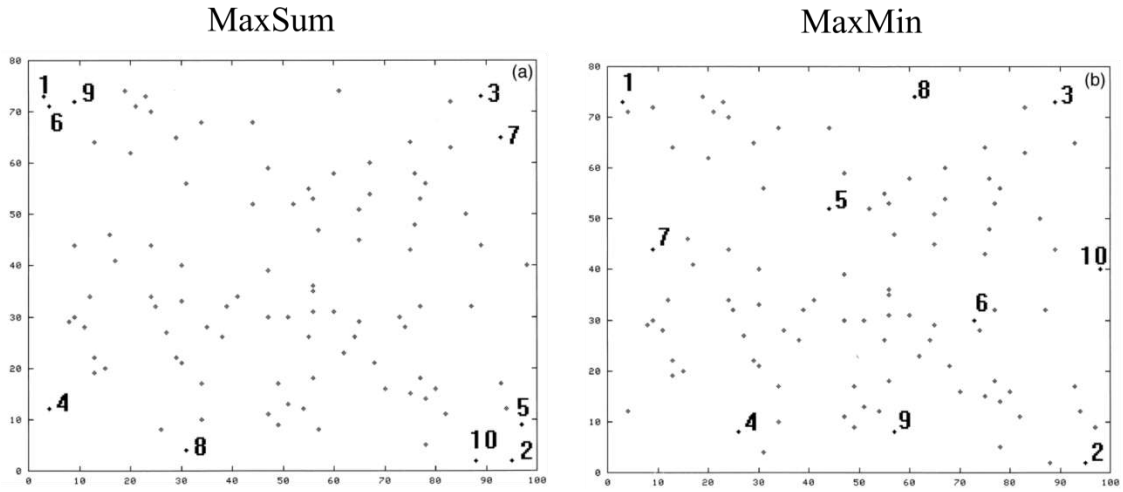


Рис. 42. Последовательность отбора 10 соединений с помощью методов MaxSum и MaxMin. Рисунок из статьи [118] публикуется с разрешения издательства. Copyright (1997) American Chemical Society.

Как уже отмечалось, отбираемый с помощью рассматриваемого алгоритма набор соединений определяется выбором первого из них, которое зачастую выбирается случайно. В этом случае первое отобранное соединение однозначно определяет выбор второго. Тем не менее, в какой бы области химического пространства не находилось первое соединение, третье из-за специфики алгоритма будет помещено в «углу» химического пространства. Для уменьшения зависимости от выбора первого соединения часто после выбора третьего соединения первые два удаляют из отобранного набора. Тогда, каким бы образом не выбиралось первое соединение, характеристики отбираемых наборов будут близкими [118].

Разработаны подходы, позволяющие в значительной мере ускорить расчеты с использованием упомянутых методов отбора. Для ускорения метода MaxMin был предложен алгоритм, позволяющий находить соединение в библиотеке, максимально удаленное от уже отобранных, не проводя полного перебора соединений [119], вычислительная сложность чего была бы равна $O(nN)$. Использование алгоритмов типа k-d деревьев позволяет снизить вычислительную сложность метода MaxMin до $O(n\log N)$ [120]. Таким образом, использование метода MaxMin является предпочтительным как с точки зрения эффективности (особенно на больших библиотеках), так и качества результатов.

3.4.2.2. Алгоритм исключенной сферы

Следующий вариант работы алгоритма DBCS включает в себя выбор некоторого порогового значения меры различия (несходства) t,

и удаление из исходной библиотеки соединений, которые отличаются от уже отобранных соединений меньше, чем на данную пороговую величину. Поскольку величина *t* по сути определяет такую гиперсферу в химическом пространстве, внутри которой соединения исключаются из рассмотрения, такой алгоритм носит название *алгоритма исключенной сферы* (англ. sphere exclusion algorithm). Алгоритмы такого типа называют иногда *алгоритмами минимального несходства* (англ. minimum dissimilarity algorithm) [118]. Метод создан в 1996 году Б. Худсоном с соавторами [121] на основе более ранней работы [122] по рациональному отбору заместителей для создания библиотеки соединений. Алгоритм работает следующим образом:

1. Задается пороговая величина несходства *t*,
2. В исходной библиотеке выбирается соединение и перемещается в набор отобранных соединений,
3. В исходной библиотеке удаляются все соединения, у которых несходство с отобранными меньше *t*,
4. Если в исходной библиотеке остались соединения, вернуться на шаг 2.

Возможно несколько вариантов этого алгоритма в зависимости от того, как выбирается соединение на шаге 2. Так, могут отбираться соединения, максимально несходные с уже отобранными. Определять меру несходства можно, например, как в методах MaxMin или MaxSum, что будет приводить к разным результатам. В оригинальном подходе Хадсона [121] на шаге 2 выбирались соединения, ближайшие к центроиду распределения отобранных соединений. Отбор соединений можно осуществлять случайным образом, что существенно ускоряет вычисления, как, например, в алгоритме MDISS [117] (часть пакета DiverseSolutions в Sybyl-X [123]), однако этот подход недетерминистский – его результаты уже однозначно не определяются выбором первого соединения. Вычислительная сложность такого алгоритма не больше $O(nN)$, хотя в реальности может быть и меньше, поскольку на каждом шагу может исключаться из рассмотрения большое количество соединений.

На Рис. 43 приведена последовательность отбора соединений с использованием алгоритма MDISS (метод исключенных сфер со случайным выбором нового соединения). Сравнивая результаты с Рис. 42, можно заметить, что этот метод дает весьма неплохие результаты, обеспечивая равномерное покрытие соединений в химическом пространстве. Следует отметить, что количество отбираемых соединений и их разнообразие существенно зависят от выбора параметра *t*.

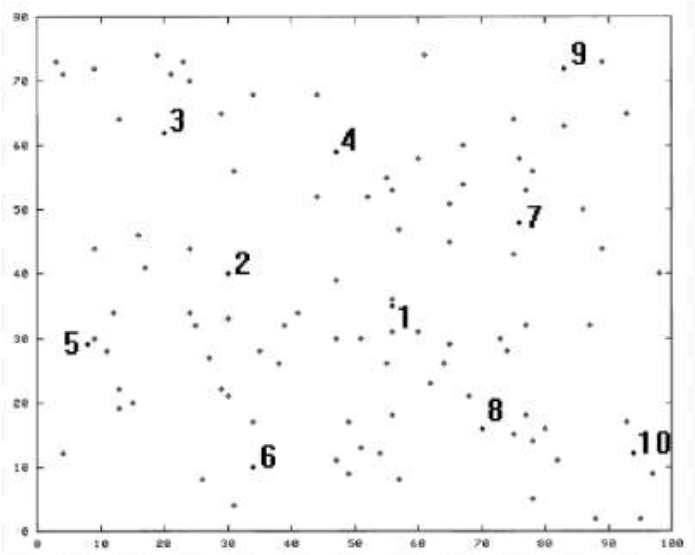


Рис. 43. Последовательность отбора 10 соединений с помощью алгоритма MDISS. Рисунок из статьи [118] публикуется с разрешения издательства. Copyright (1997) American Chemical Society.

3.4.2.3. Алгоритм OptiSim

Дж. Холлидей и П. Виллетт в своей работе по методам максимального несходства [116] указали, что все они (за исключением MaxMin) могут быть существенно улучшены за счет введения дополнительного правила отбора: каждый раз отбирается то соединение, которое отличается от любого уже отобранного больше, чем на заданную величину R_{min}. Это правило роднит этот алгоритм с методами минимального несходства (например, исключенной сферы). Р. Кларк предложил алгоритм оптимизируемого отбора из K-несходных (англ. Optimizable K-Dissimilarity Selection) или OptiSim [118], для которого методы максимального и минимального несходства являются частными случаями.

Алгоритм OptiSim заключается в следующем. Создается четыре списка: конечный набор, промежуточный набор, список кандидатов и корзина. Все они на первом шаге берутся пустыми. Полагается, что задачей является включение в конечный набор *M* наиболее разнообразных соединений. Далее выполняется следующее:

1. Выбрать случайным образом одно соединение из исходной библиотеки из *N* соединений и поместить его в конечный набор. Все остальные *N-1* соединения поместить в список кандидатов.

2. Из списка кандидатов выбрать случайным образом соединение, которое отличаются от соединений из конечного набора больше, чем на заданную величину R_{min}, и поместить его в промежуточный набор.
3. Повторять шаг 2, пока в промежуточном наборе не окажется K соединений.
4. Если в промежуточном наборе меньше K соединений и список кандидатов опустел, переместить соединения из корзины в список кандидатов и перейти на шаг 2.
5. Если после пройденных шагов промежуточный набор пуст, то завершить работу алгоритма и выйти, выдав в качестве результата сформированный конечный набор.
6. Найти в промежуточном наборе соединение, которое имеет наибольшую меру несходства с соединениями из конечного набора, и переместить его в последний. Оставшиеся соединения из промежуточного набора переместить в корзину.
7. Если в конечный набор выбрано требуемое число соединений M, остановить алгоритм и выдать список соединений из конечного набора. Иначе перейти на шаг 2.

Первые три шага работы алгоритма приведены на Рис. 44.

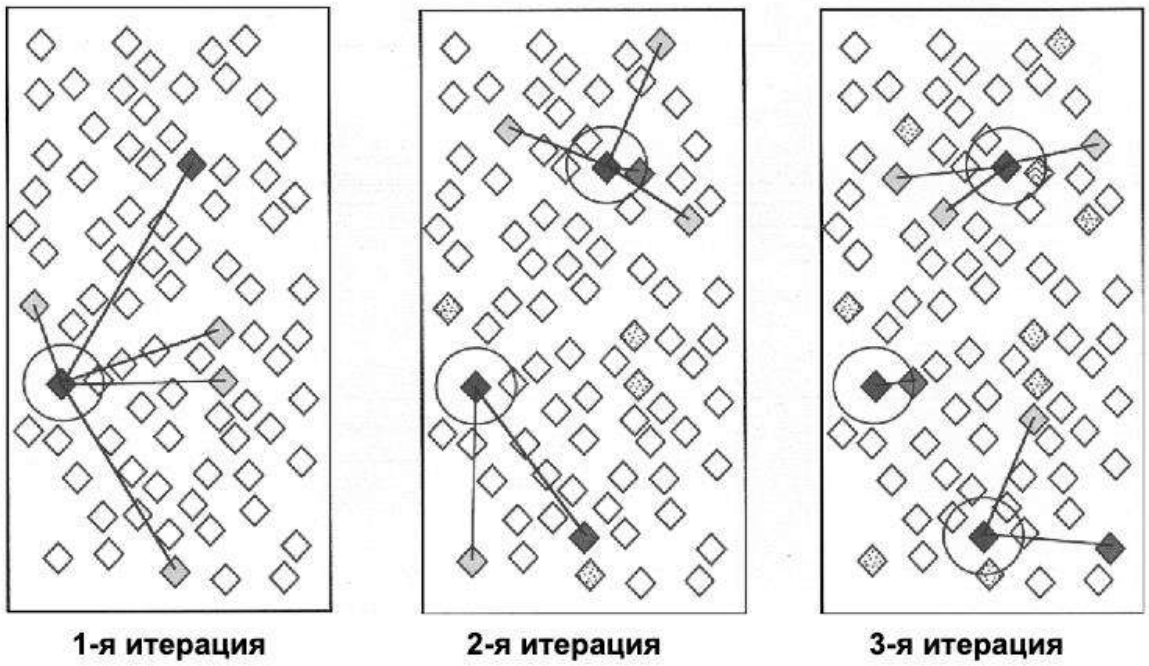


Рис. 44. Первые три итерации отбора соединений методом OptiSim. Рисунок из статьи [118] публикуется с разрешения издательства. Copyright (1997) American Chemical Society.

Можно заметить, что алгоритм исключенной сферы является частным случаем алгоритма OptiSim при $K = 1$. Методы максимального несходства являются его частыми случаями при $K = N$. Подбором числа K можно добиться баланса между предельным разнообразием, которое обеспечивается методами максимального несходства, и покрытием всех областей химического пространства, что является достоинством методов минимального несходства.

3.4.3. Отбор соединений на основании разбиения химического пространства

Кластеризацию и методы DBCS иногда называют подходами, основанными на расстоянии, поскольку их работа основана на использовании матрицы расстояний между химическими соединениями в химическом пространстве. Эти методы разделяют соединения на группы путем вычисления расстояний между ними. Альтернативой этому является разделение самого химического пространства на участки и отбор соединений из каждого (или некоторых) из них. Наиболее распространенный способ такого разделения основан на задании в химическом пространстве небольшого числа ортогональных осей и разделении каждой из них на набор «корзин», Рис. 45. Корзины не обязательно должны быть одинаковыми – некоторые из них могут покрывать бесконечные интервалы (например, $\log P < 0$). Это деление выделяет в химическом пространстве некоторое число ячеек (англ. *cell*), что дало название самому подходу – *ячеечные методы отбора* (англ. *cell-based methods*). Существуют ячеечные методы, в которых каждая ячейка может образовываться не пересечением плоскостей, параллельных координатным плоскостям (Рис. 45a), а иметь собственные размеры (Рис. 45b). Формирование набора несходных соединений с использованием ячеечных подходов включает две стадии: размещение соединений по ячейкам (аналог кластеризации) и отбор одного или нескольких соединений в формируемый набор.

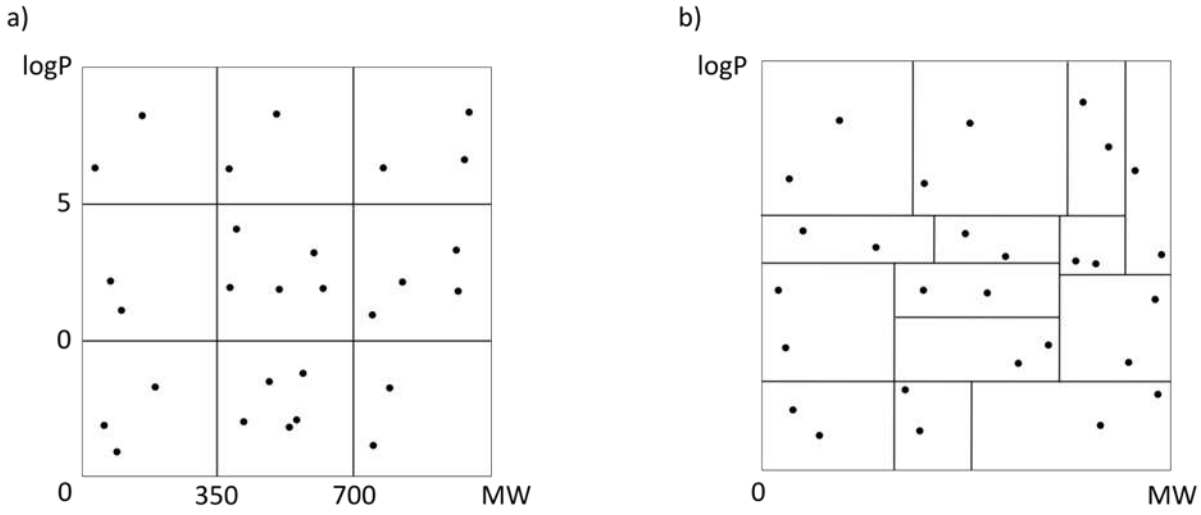


Рис. 45. Разбиение соединений некоторой базы химическом пространстве, осями которого являются значения молекулярной массы (MW) и липофильности (log*P*). Слева (а) – ячейки образуются пересечениями линий, параллельных осям координат, справа (б) – размер каждой ячейки задается индивидуально.

Основной проблемой данного подхода является крайне быстрый рост числа ячеек с количеством осей N и корзин M, выделяемых на каждой из них: если все оси имеют одинаковое число корзин, то число ячеек равно M^N. При разном числе корзин число ячеек равно $\prod_{i=1}^{N} M_i$, где M_i – число корзин на i-й оси. Поэтому для проведения анализа необходимо использовать минимально возможное число осей. Поэтому для этой цели в качестве осей координат в химическом пространстве неэффективно использовать фингерпринты и структурные ключи. В этом случае наиболее эффективно использовать в качестве осей ортогональные дескрипторы или физико-химические свойства соединений (logP, MW, число H-акцепторов).

3.4.3.1. Разделение химического пространства дескрипторов

В одной из первых работ по использованию ячеечного подхода к созданию диверсифицированного набора соединений [124] Льюисом с соавторами проводился статистический анализ дескрипторов, из которых было выбрано 6 почти ортогональных дескрипторов, представляющих различные типы физико-химических свойств: гидрофобность (ClogP), полярность (нормализованная сумма квадратов атомных электротопологических индексов), способность к взаимодействиям между ароматическими кольцами (плотность

ароматичности – число ароматических колец, деленное на молекулярную массу), способность к образованию водородных связей (число Н-акцепторов и Н-доноров), форму молекулы (индекс гибкости – произведение $^{1}к$ и $^{2}к$ индексов Кьера-Холла, деленное на число вершин). Каждая ось делилась на 2-4 корзины с получением 576 ячеек, в которых было размещено 150 000 соединений. Из каждой ячейки отбирались соединения для создания небольшого набора разнообразных соединений.

Многомерное пространство дескрипторов может быть приведено к низкоразмерному с помощью метода главных компонент, который определяет в химическом пространстве набор из малого количества ортогональных осей (главных компонент), соответствующих линейным комбинациям исходных дескрипторов (см. раздел 2.3.2 в пособии 4). Поскольку получаемые в результате такого анализа главные компоненты часто не имеют четкой физико-химической интерпретации, в факторном анализе осуществляют вращение ортогональных осей с целью придания получающимся при этом дескрипторам определенного смысла. Так, например, в работе [125] было показано, что можно найти 4 дескриптора, объясняющих 90% разброса данных (дисперсии), и 7 дескрипторов, которые объясняют 95% дисперсии.

Ручное или автоматическое определение ячеек химического пространства, независимых в общем случае от того, как молекулы расположены, заменяет проблему группирования химических соединений проблемой подбора наиболее адекватного способа выделения ячеек в химическом пространстве. Выбор числа и интервалов для корзин на осях и вообще определение ячеек химического пространства (то есть биннинг химического пространства) является одной из ключевых задач для отбора соединений этим методом. Количество корзин на различных осях может как совпадать, так и отличаться. Существует несколько способов биннинга [126]: в *биннинге, независимом от дескрипторов,* выделение корзин на одной координатной оси не зависит от того, как выделяются корзины на остальных осях. В *биннинге, зависимом от дескрипторов,* положение гиперплоскости в *D*-ом измерении химического пространства зависит от того, как проведены гиперплоскости в оставшихся *D-1* измерениях, в результате чего ячейки «подгоняются» под определенные параметры. Контролировать распределение ячеек можно с помощью различных параметров. В *биннинге, зависимом от интервалов,* размеры ячеек стараются сделать одинаковыми (например, делением на равные участки интервала

разброса значений каждого из используемых дескрипторов для библиотеки соединений). В *биннинге, зависимом от заселенности,* размеры ячеек определяются количеством соединений, попадающих в ячейки, что приводит к равномерной заселенности ячеек. Было показано [126], что биннинг, обеспечивающий одинаковую заселенность ячеек, позволяет более эффективно вести последующий отбор несходных соединений.

Главной особенностью ячеечных методов отбора соединений является отсутствие необходимости расчета попарных расстояний между ними. Это обеспечивает вычислительную сложность метода от $O(1)$ (отсутствие зависимости от числа соединений в исходной базе), до $O(N)$ (пропорционально числу соединений в базе), что позволяет что позволяет очень быстро вести отбор несходных соединений в базах данных даже очень большого размера. Существуют и другие преимущества такого подхода. Ячеечные методы отбора позволяют легко идентифицировать пустые ячейки или ячейки с малой заселенностью, что позволяет выделить области химического пространства, плохо предоставленные в базе данных. С другой стороны, возникновение пустых областей может быть обусловлено определенной взаимозависимостью дескрипторов, то есть их неортогональностью. К примеру, если в молекуле присутствует большое число Н-акцепторов или Н-доноров, то сомнительно, что данная молекула будет обладать высокой гидрофобностью ($\log P$).

Анализ распределения соединений по ячейкам позволяет легко оценивать диверсифицированность отбираемых наборов соединений и таким образом проводить их сравнение. Простейшим методом оценки диверсифицированности набора соединений является подсчет количества ячеек, занимаемых соединениями отобранного набора в разделенном на ячейки химическом пространстве. Другими мерами диверсифицированности набора соединений (как всей базы, так и отобранного набора) являются статистические критерии, связанные с χ^2-распределением, в частности, критерий Пирсона:

$$\chi^2 = \frac{\sum_{i=1}^{n}\left(N_i - \frac{N}{n}\right)^2}{N/n}, \qquad (40)$$

равный 0 для равномерного распределения по ячейкам (то есть предельно разнообразного), или критерий, использованный Уолдманом [112]:

$$D_{\chi^2} = -\sum_{i=1}^{n}\left(N_i - \frac{N}{n}\right)^2, \qquad (41)$$

где в обоих случаях N_i – число соединений, попавших в i-ю ячейку, N – общее число соединений в отобранном наборе, n – общее число ячеек. Из-за наличия в формуле знака «минус», чем больше разнообразие набора (чем более равномерно распределены соединения по ячейкам), тем значение D_{χ^2} выше. С помощью этих критериев можно оценивать диверсифицированность (разнообразие) наборов или оптимизировать выделение ячеек. Недостатком использования этих статистических критериев является то, что для них равномерное заполнение ячеек является более предпочтительным, чем заполнение большего числа ячеек. Они не удовлетворяют первым трем требованиям Уолдмэна и поэтому, например, для них, при распределении 6 соединений по 3 ячейкам, распределения (3,3,0) и (4,1,1) являются эквивалентными. Еще одной характеристикой, по которой можно оценивать равномерность распределения соединений химическом пространстве с использованием ячеечных методов, является энтропия:

$$S = -\sum_{i=1}^{n} \frac{N_i}{N} \ln\left(\frac{N_i}{N}\right). \qquad (42)$$

Энтропия уже лишена указанного недостатка рассмотренных выше критериев – при ее использовании распределение (4,1,1) будет более предпочтительным, чем (3,3,0). Однако она все же не удовлетворяет первым двум требованиям Уолдмэна.

3.4.3.2. Разделение с использованием фармакофорных ключей

Вместо классических дескрипторов для размещения соединений в химическом пространстве могут использоваться трех- и четырехточечные фармакофоры (см. ниже раздел 4). Особенно широко используются для этого трехмерные фармакофоры. Полагается, что каждый возможный фармакофор обозначает одну ячейку. Тогда соединения могут быть размещены в таком химическом пространстве на основании фармакофоров, которые в них содержатся. В отличие от разделенного на ячейки дескрипторного пространства, каждое соединение в таком фармакофорном «пространстве» может занимать одновременно несколько ячеек. Таким образом, задача отбора соединений в данном случае сводится к отбору соединений, обладающих в совокупности наибольшим разнообразием фармакофоров.

3.4.4. Методы оптимизации диверсифицированных наборов соединений

Как уже отмечалось, задача поиска оптимального диверсифицированного набора соединений (то есть набора максимально несходных между собой соединений) относится к классу NP-проблем. До этого мы обсуждали приближенные методы ее решения, основанные на определении занятых областей химического пространства и итеративного выбора по одному (или несколько) соединений из каждой такой зоны. Как часто бывает, комбинаторные проблемы могут решаться с использованием стохастических (то есть основанных на использовании генераторов случайных чисел) методов оптимизации.

3.4.4.1. Оптимизация набора соединений с помощью стохастических алгоритмов

Методы оптимизации наборов структур имеют целью проводить поиск оптимального (диверсифицированного) поднабора с использованием стохастических методов. В отличие от вышеуказанных методов отбора соединений, в рамках данного подхода первоначальный отбор соединений является лишь первым этапом работы, а на следующем этапе с помощью стохастических алгоритмов ведется формирование наборов, оптимизирующих функцию, характеризующую диверсифицированность наборов соединений. Для осуществления этого могут использоваться любые методы стохастической оптимизации, предназначенные для оптимизации функций нескольких переменных. Таким образом, для проведения отбора необходимо: (а) определить вид целевой функции, подлежащей оптимизации при формировании оптимального набора структур, (б) определить, каким образом проводить операции над векторами оптимизируемых параметров.

Обычно для отбора соединений используется метод Монте-Карло в сочетании с процедурой «*искусственного отжига*» (англ. simulated annealing) [127]. В методе Хассана с соавторами [128] в качестве функции разнообразия используется минимальный квадрат расстояния между молекулами, величины

$$\left(\sum \left(d_{ij}^2\right)^m / N_d\right)^{1/m}$$

и

$$\left(\prod d_{ij}^{2}\right)^{1/N_d}$$

где d_{ij} – расстояние между соединениями i и j, N_d – число расстояний между соединениями, m – некоторое заранее определенное целое число, причем в приведенных выше формулах суммирование и перемножение ведется по всем парам химических соединений. В качестве строки оптимизируемых значений используется номера молекул, вошедших в конечный набор. На первом этапе строка инициируется случайными номерами соединений из общего набора. На следующем шагу случайно выбранный элемент строки заменяется случайно выбранной молекулой из общего набора и вычисляется изменение функции разнообразия (обратной целевой функции). Заметим, что выбор целевой функции во многом определяет качество отобранного набора.

Отбор соединений производится с помощью критерия Метрополиса: если после замены соединения набор становится более разнообразным, то он используется для следующей итерации («принимается»), если он становится менее разнообразным, то вероятность того, что он будет принят, зависит от величины $\exp[-\Delta D/kT]$, где ΔD – разница между значениями функции разнообразия нового и исходного набора. Процедура оптимизации останавливается по достижении определенного числа шагов или после того, как значение функции разнообразии не улучшилось после заданного числа шагов.

При использовании метода Монте-Карло в сочетании с методом «искусственного отжига» [127] по мере понижения «температуры» T в критерии Метрополиса отбираемые наборы соединений постепенно локализуется вблизи оптимума, а при нулевой «температуре» разрешены только такие изменения, которые приводят к увеличению разнообразия, и таким образом находится локально оптимальный набор. Чтобы найти глобально оптимальный набор процедуру «искусственного отжига» повторяют несколько раз, и из найденных наборов отбирают наилучший. При увеличении числа повторений вероятность найти истинный глобальный минимум повышается.

Интересный подход к оптимизации наборов соединений был предложен в работе Д. Аграфиотиса [129]. В нем в качестве целевой функции используется статистический критерий Колмогорова-Смирнова (статистика Колмогорова-Смирнова) для сравнения

интегральной функции распределения[1] значений расстояний между отобранными соединениями $F(x)$ с заданной интегральной функцией распределением вероятности $F^*(x)$. Критерий Колмогорова-Смирнова равен максимальному абсолютному отклонению эмпирической функции распределения от заданной (Рис. 46):

$$K = \max_{-\infty<x<\infty} |F(x) - F^*(x)|. \quad (43)$$

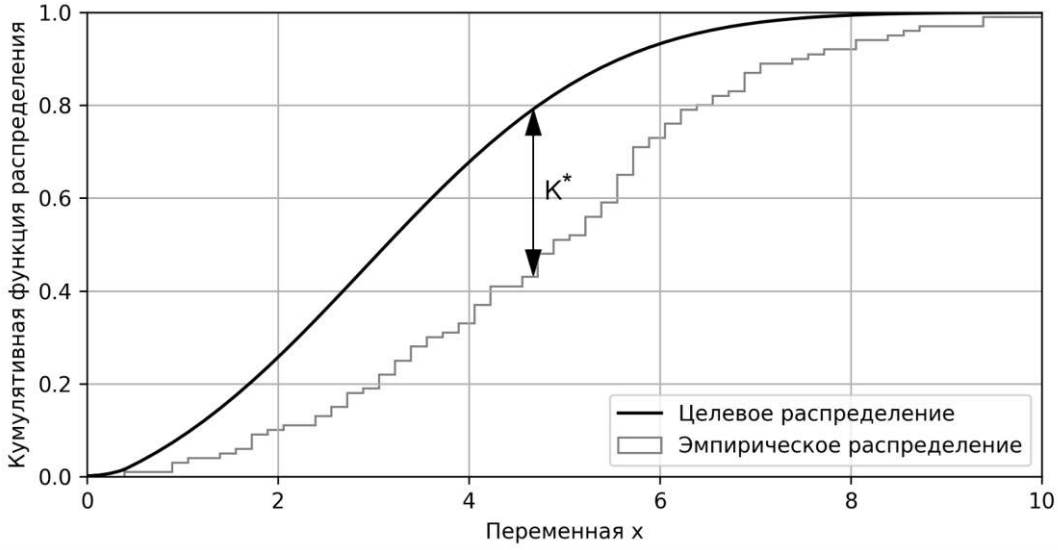


Рис. 46. Использование критерия Колмогорова-Смирнова для нахождения отличий эмпирической функции распределения от заданного.

Значения этого критерия легко рассчитать – достаточно отсортировать расстояния между соединениями и за один проход по отсортированному списку определить максимальное отклонение от заданного распределения. В отличие от χ^2-статистики, описанной выше, его расчет не требует биннинга химического пространства. Использование критерия Колмогорова-Смирнова в создании диверсифицированных наборов базируется на том, что для

[1] Функция распределения $\Phi(x)$ (называемая также интегральной функцией распределения, англ. cumulative или integral distribution function) – это функция, которая равна вероятности того, что некоторая (случайная) величина примет значение от -∞ до x. Не следует ее путать с плотностью распределения вероятности $f(x)$, которая есть производная функции распределения $f(x) = d\Phi(x)/dx$. В случае, например, нормального распределения функция Гаусса характеризует плотность вероятности распределения, а функция ошибок – интегральную функцию распределения (с точностью до шкалирования и сдвига $\Phi(x) = 0.5[1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})]$).

равномерного распределения соединений в химическом пространстве достаточно большой размерности распределение расстояний между ними стремится к нормальному. Параметры этого распределения (среднее и дисперсия) могут быть легко определены, например, с помощью модельного распределения. Для этого создается случайное равномерное распределение, покрывающее то же химическое пространство, что и база, из которой происходит отбор, и определяется среднее расстояние и дисперсия в нем. Если метрика в пространстве задается манхэттеновским расстоянием, то среднее и дисперсия вычисляются в аналитическом виде [129]. Далее методом «искусственного отжига» отбираются соединения, которые обеспечивают распределение, максимально близкое к заданному, то есть равномерному.

Преимущество использования методов оптимизации заключается в том, что сложность алгоритма зависит от числа химических соединений только в отбираемой выборке, которая, как правило, много меньше выборки соединений, из которых происходит отбор. Таким способом можно проводить отбор из баз, содержащих миллионы соединений, что недоступно большинству методов кластеризации. Кроме того, этот метод потенциально позволяет находить наилучшие (глобально оптимальные) решения – то есть формировать наборы соединений, максимально диверсифицированные по заданному критерию разнообразия.

3.4.4.2. Многоцелевая оптимизация набора соединений

Во всех рассмотренных выше методах отбора соединений алгоритм руководствовался только одной целью – найти наиболее разнообразный набор соединений. Существуют также методы стохастической оптимизации, которые позволяют в процедуре отбора руководствоваться одновременно несколькими целями. Например, очень желательно, чтобы при формировании диверсифицированного набора принималась во внимание цена соединений, поэтому надо одновременно принимать во внимание два фактора – разнообразие набора и общую стоимость входящих в него соединений. Для этого информация об оптимизируемых факторах должны быть внесена в целевую функцию, которая потом оптимизируется (чаще всего – минимизируется) с помощью рассмотренных выше стохастических методов.

Существует 2 основных подхода к многоцелевой оптимизации: метод взвещшенных сумм и парето-оптимизации.

Метод взвешенных сумм

В методе взвешенных сумм целевая функция принимается равной сумме произведений факторов, которые необходимо оптимизировать, на веса, характеризующие важности этих факторов. Факторы, которые необходимо минимизировать, входят в нее с положительными весами, которые необходимо максимизировать – с отрицательными. Например, если необходимо максимизировать разнообразие набора D и минимизировать его цену P, тогда целевую функцию можно представить следующим образом:

$$F = -w_1 D + w_2 P \qquad (44)$$

Существует много способов проведения многоцелевой оптимизации, многие из которых используются в комбинаторной химии [130-132]. Ниже описаны те из них, которые чаще всего применяются для формирования наборов химических соединений на примере программы DirectedDiversity [132].

В программе DirectedDiversity [132] пользователь может комбинировать следующие функции:

- функцию сходства с заданными соединениями – обычно среднее расстояние от набора из M отобранных соединений до ближайшего к нему набора из L предопределенных лидеров (соединений, на которые отобранный набор должен быть похож):

$$S = \frac{1}{M}\sum_{i=1}^{M}\min_{j=1}^{L}(d_{ij}) \qquad (45)$$

Эту функцию обычно необходимо минимизировать.

- функцию, оценивающую разнообразие отобранного набора – обычно среднее расстояние от каждого из соединений набора до его ближайшего соседа в наборе:

$$D = \frac{1}{M}\sum_{i=1}^{M}\min_{j\neq i}^{M}(d_{ij}) \qquad (46)$$

Эту функцию обычно максимизируют.

- функцию, характеризующую дополнительность набора – критерий, нужный для заполнения пустых областей в указанной коллекции S^*, состоящей из M^* соединений (молекул). Отобранные соединения (набора S) должны быть не похожи на соединения из коллекции S^* и при этом не похожи друг на друга. Поэтому этот критерий есть ничто иное, как разнообразие объединенного набора из отобранных соединений и предопределенной коллекции:

$$D(S,S^*) = \frac{1}{M}\sum_{i=1}^{M+M^*} \min_{j\neq i}^{M+M^*}(d_{ij}) \qquad (47)$$

Этот критерий необходимо максимизировать.

- ограничения по значениям свойств – характеризуют отклонения в свойствах соединений из отобранного набора от заданного интервала значений и равно среднему по набору штрафов за выход значений свойств (всего H свойств, на которые накладываются ограничения) из заданного интервала. Если j-е свойство i-го соединения x_{ij} находится в пределах интервала $[x_j^{min}, x_j^{max}]$, то штраф полагается равным 0, в противном случае он равен величине отклонения от минимального или максимального значения интервала:

$$P = \frac{1}{M}\sum_{i=1}^{M}\sum_{j}^{H} \max(x_j^{\min} - x_{ij}, x_{ij} - x_j^{\max}, 0) \qquad (48)$$

Этот критерий необходимо минимизировать.

- распределение характеристик соединений – насколько совпадает эмпирическая функция распределения $F(x)$ выбранной характеристики с заданным распределением $F^*(x)$. Штрафом за несовпадение является значение описанной выше статистики Колмогорова-Смирнова. Такой характеристикой может быть, например, молекулярный вес, липофильность, расстояние между соединениями в наборе (тогда эта величина будет характеризовать разнообразие), а также другие заданные пользователями характеристики, а заданным распределением – нормальное, равномерное или распределение в заданной базе данных. Этот критерий минимизируется.

- функция, характеризующая свойства (активность) соединений – среднее значение свойства (активности) соединений в отобранном наборе по результатам предсказаний с помощью построенных моделей QSAR/QSPR. При рассмотрении биологической активности эту функцию обычно максимизируют.

- селективность – функция, характеризующая селективность по отношению к набору биологических мишеней, то есть насколько хорошо соединения действуют на заданную мишень q и плохо – на другие мишени j (общее число таких мишеней равно G):

$$Q = \frac{1}{M}\sum_{i=1}^{M}\left[a_{iq} - \max_{j\neq q}^{G}(a_{ij})\right] \qquad (49)$$

Обозначения аналогичны приведенным выше. Поскольку с ростом Q селективность возрастает, эту функцию обычно максимизируют.

- перекрывание – критерий, который измеряет перекрывание отобранного набора соединений S с другим шаблонным набором S^*:

$$O(S,S^*) = 1 - \frac{N_{S \cap S^*}}{N_S} \qquad (50)$$

где $N_{S \cap S^*}$ - число соединений, присутствующих одновременно в наборах S и S^*, N_S - только в отобранном наборе S. Обычно требуется, чтобы соединения не повторялись, поэтому этот критерий максимизируется.

Метод взвешенных сумм сводит проблему многоцелевой оптимизации к одноцелевой оптимизации целевой функции. Одной из главных проблем использования многоцелевой оптимизации такого типа является необходимость задавать значения весов, характеризующих важность каждого из факторов, тогда как не существует каких-либо универсальных или обоснованных способов выбора значений таких весов. Поэтому их обычно выбирают методом проб и ошибок: из нескольких вариантов выбирают тот, который дает наилучшие с точки зрения пользователя результаты.

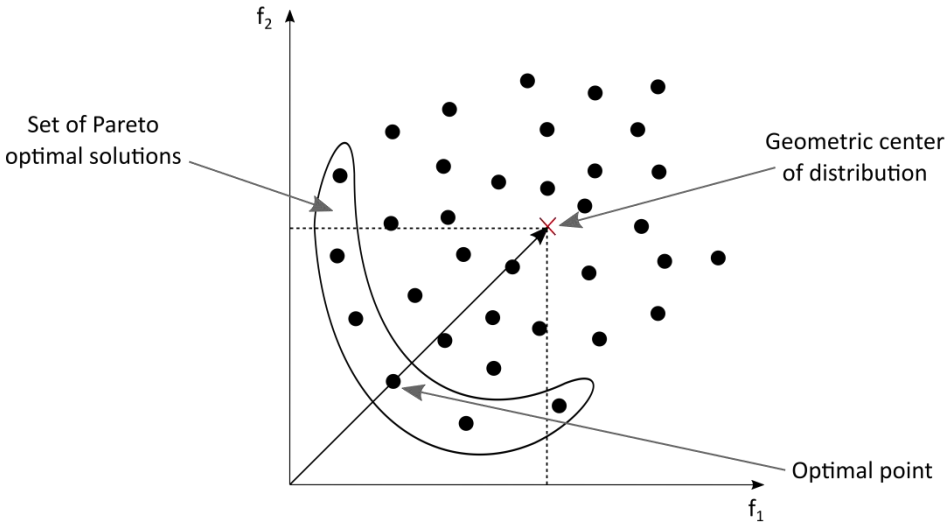


Рис. 47. Выбор множества Парето-оптимальных решений и выбор лучшего из них.

Существует, однако, и другой способ проведения многоцелевой оптимизации – с использованием принципа Парето. Этот принцип пришел в многокритериальную оптимизацию из экономики, и

сформулирован В. Парето следующим образом: «Всякое изменение, которое никому не приносит убытков, а некоторым людям приносит пользу (по их собственной оценке), является улучшением». Идея метода заключается в следующем: предположим, что с помощью отбора соединений необходимо минимизировать D функций штрафа, зависящих от выбранного набора. Очевидно, что наилучшее решение – это то, которое минимизирует одновременно все функции штрафа. Построим график, координатными осями которого являются значения всех D функций штрафа (графически его можно построить только когда $D = 2$, однако программно можно реализовать этот метод для любого D), координатами точек на котором являются значения функций штрафа для отобранных наборов, Рис. 47. Множество Парето-оптимальных решений – это набор таких решений – точек на данном графике, из которых нельзя перейти в другую точку графика, улучшая (в нашем случае уменьшая) значения всех оптимизируемых функций штрафа. Эти точки образуют на графике линию (в 3-мерном пространстве – поверхность, а в произвольном многомерном пространстве – гиперповерхность). Переход из любой точки этой поверхности в другую приводит к уменьшению одних функций штрафа и увеличению других. Все остальные точки – перебранные решения проблемы – находятся по одну сторону этой поверхности. Для Парето-неоптимальных решений существует хотя бы одна точка на поверхности Парето-оптимальных решений, в которую можно перейти из данной, улучшив (уменьшив) все функции штрафа.

Принцип Парето не дает ответа, каким образом выбирать лучшее решение из множества Парето-оптимальных. Имея набор решений, в которых представлены различные компромиссы между различными целями оптимизации, создатель библиотеки может сделать обоснованный выбор единственного оптимального решения. Заметим, что в методе взвешенных сумм пользователь, придавая разные веса функциям штрафа, сам вынужден искать данный компромисс. Однако в ряде случаев может потребоваться автоматический выбор решения. Обычно для отбора лучшего решения из множества Парето-оптимальных используют:

- минимизацию линейной комбинации функций штрафа с весами, заданными вручную (аналогично изложенному выше подходу взвешенных сумм),
- геометрические соображения, например, выбирается решение, максимально близкое к линии, соединяющей начало координат и геометрический центр распределения точек на

графике (т.е. с точкой, координаты которой являются средними по выборке) [133], Рис. 47.

Метод Парето используется во многоцелевой генетической оптимизации, реализованной в программе MoSELECT [134]), в ходе которой с помощью генетического алгоритма генерируется набор различных решений. Далее т.н. недоминирующим решениям, для которых минимизируется количество точек, попадающих внутрь прямоугольника, ограниченного значениями оптимизируемых параметров (т.е. число таких точек, из которых можно без увеличения ни одного из параметров перейти к заданной) присваивается меньший ранг. Доминирующим решениям, то есть тем, которые находятся над другими точками, присваивается больший ранг (Рис. 48). Решения с большим рангом имеют меньшую вероятность к скрещиванию (то есть реже вовлекаются в кроссинговер и, следовательно, оставляют меньшую информацию о себе в популяции), и могут быть выброшены при генерации новой популяции. Таким образом, через некоторое число шагов генетической оптимизации пользователь будет иметь набор недоминирующих, Парето-оптимальных решений. Зная распределение решений с различными компромиссами штрафов, пользователю не требуется, как в подходе взвешенных сумм, перебирать различные наборы весов (то есть варианты компромиссов между различными штрафами), чтобы выбрать оптимальный набор – множество возможных вариантов предоставляется ему программой.

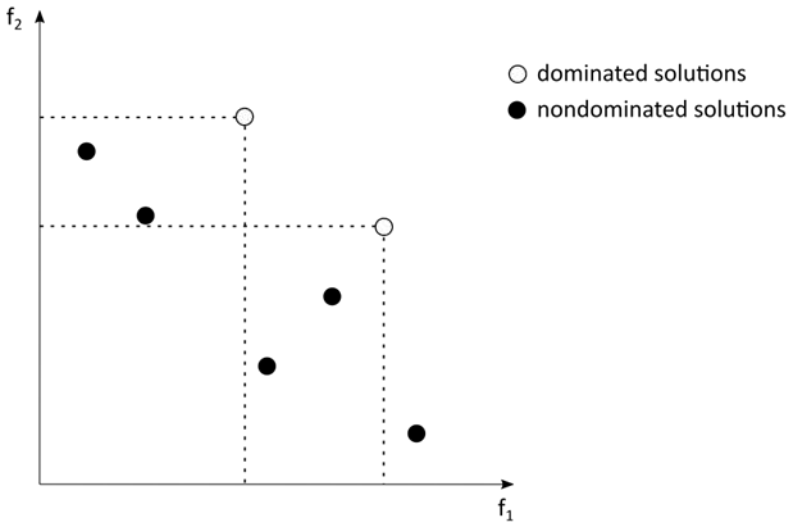


Рис. 48. Ранжирование различных решений в методе двуцелевой генетической оптимизации.

3.5. ТЕОРЕТИЧЕСКАЯ КОМБИНАТОРНАЯ ХИМИЯ

Экспериментальная комбинаторная химия была создана в конце 1980х – начале 1990х годов [135] и привела к появлению большого числа автоматических методов для химического синтеза [136, 137]. Общей особенностью данных методов является возможность проведения синтеза одновременно множества различных соединений. Хотя появление и развитие комбинаторной химии было вызвано нуждами фармацевтической отрасли, в настоящее время существенный интерес к ней проявляют и в других областях химии – например, дизайне материалов [138, 139].

В отличие от классического способа синтеза одного соединения за другим, главная идея комбинаторного химии заключается в одновременном синтезе всех соединений библиотеки путем комбинирования реагентов – *строительных блоков* (англ. building block), что и дало название этому подходу, Рис. 49.

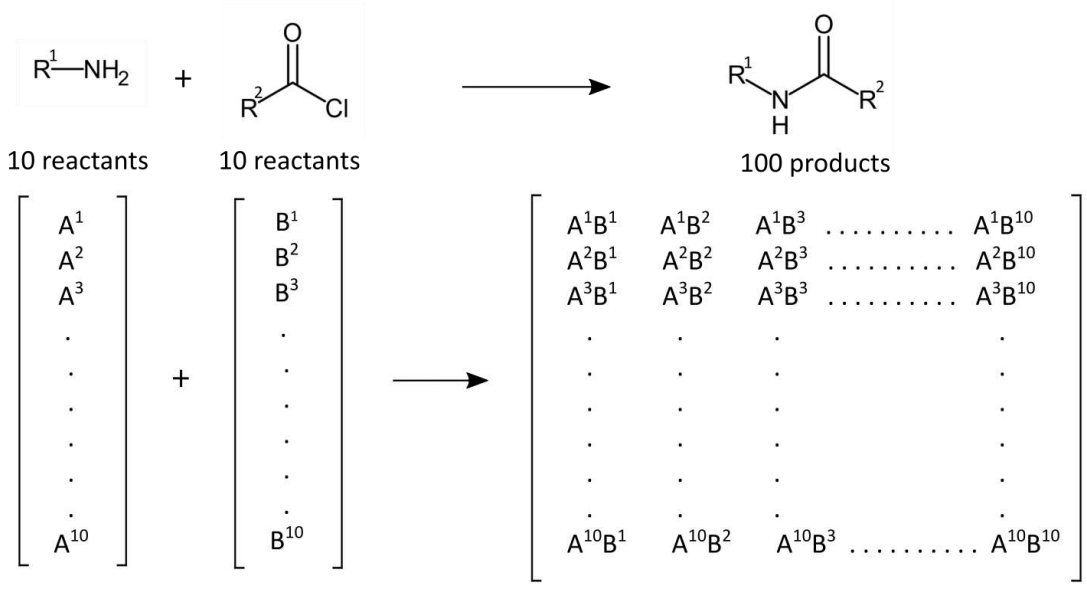


Рис. 49. Общая идея комбинаторного синтеза. Каждый из реагентов одного типа реагирует с реагентами другого.

Каждый из строительных блоков состоит из заместителя (называемого R-группой), который переходит в конечное соединение, и реакционного центра, который участвует в реакционном превращении и образует некоторый каркас, общий для всех соединений комбинаторной библиотеки. Таким образом, комбинаторная химия позволяет получить конгенерный ряд соединений (то есть с общим остовом). Если мы имеем 10 реагентов

одного типа и 10 – другого, которые вступают в реакцию попарно, то можно получить 100 соединений.

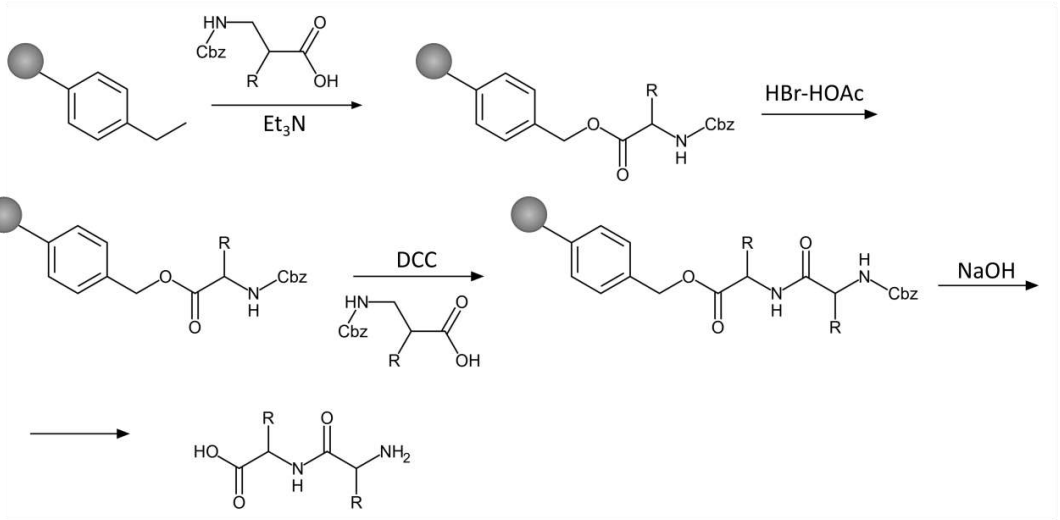


Рис. 50. Схема синтеза полипептидов с использованием твердой подложки (CBZ – бензилоксикарбонильная защитная группа, DCC - дициклогексилкарбодиимид)

Еще большее число соединений можно получать, если реакцию комбинаторного синтеза проводить несколько раз. Обычно для этого используется техника комбинаторного синтеза на поверхности твердого носителя, состоящая в том, что на поверхность полимерного носителя пришивается один из реагентов. Далее на эти зерна полимера (англ. bead) действуют большим избытком другого реагента, чтобы реакция прошла до конца. Избыток реагента легко удаляется фильтрованием и промывкой. Эта процедура наиболее широко используется в синтезе пептидов (Рис. 50). В настоящее время существуют автоматические синтезаторы, позволяющие проводить синтез пептидов программируемого строения.

Использование процедуры *разделения-смешения* (англ. split-mix) в сочетании с синтезом на поверхности твердого носителя позволяет легко получать очень большое число соединений. Рассмотрим на примере пептидного синтеза, для которого эта процедура впервые и была использована. Предположим, мы хотим получить все возможные полипептиды, состоящие из трех типов аминокислот. Для этого на полимерный носитель ковалентно пришивается три аминокислоты. Затем носитель с пришитыми аминокислотами разделяется на три части, а на каждую из них опять действуют тремя разными аминокислотами (Рис. 51). В итоге получается 3·3 = 9 различных

дипептидов, которые опять смешивают. Разбивая каждую из них на 3 части и смешивая, можно получить после проведения реакции 27 различных продуктов. Таким образом, количество продуктов экспоненциально растет с числом повторений.

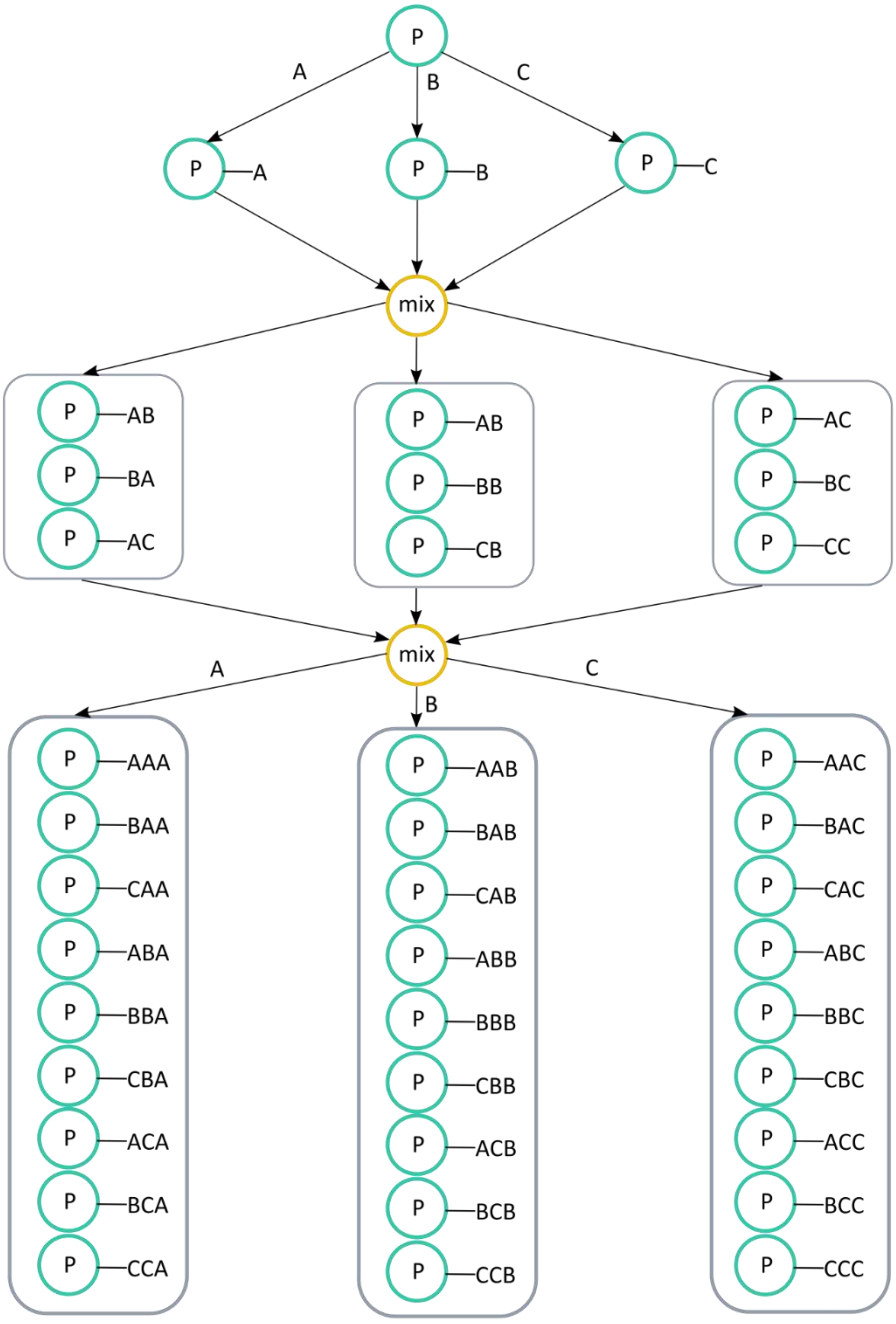


Рис. 51. Комбинаторный синтез белков из 3х аминокислот (A, B и C) с использованием процедуры разделения-смешения

Быстрого роста числа соединений в комбинаторной библиотеке можно добиваться также с помощью последовательной *дериватизации* (т.е. получения нового соединения из предшественника) исходного

темплата (остова), который содержит несколько реакционных центров, реагирующих с различными типами строительных блоков.

Наиболее широко комбинаторные библиотеки используются для биологического скрининга. Для проведения скрининга используются специальные установки. Основная часть ее – плашки (англ. plates), в которых имеются лунки для помещения испытуемых веществ. В лунки добавляются также испытуемая *биологическая система* (англ. bioassay), которая часто представляет собой белок со связанным внутри него лигандом. Если испытуемое вещество обладает высоким сродством к белку, оно высвобождает лиганд, что вызывает некоторый физический отклик (люминесценцию, поглощение света определенной длины волны и др.) свободным лигандом. Такие соединения принято называть *хитами* (англ. hit). Существует множество различных биологических испытаний, которые позволяют определять различные качественные и количественные характеристики веществ, в том числе фармакокинетические параметры, такие как проникновение через гематоэнцефалический барьер, распределение вода-октанол, кровь-воздух и др.

Поскольку в ходе высокопроизводительного скрининга за одну кампанию могут быть проверены миллионы соединений, проблема хранения и доставки библиотек для скрининга начинает представлять серьезную проблему. Действительно достаточно сложно организовать добавление миллиона различных соединений в лунки плашек. В то же время комбинаторная химия позволяет решить эту проблему достаточно эффективно – необходимо иметь относительно небольшое число реагентов для того, чтобы получить большое число новых соединений. Для проведения биологического скрининга возможно создание комбинаторной библиотеки двух типов. Во-первых, в каждой лунке плашки можно получать соединения определенного типа добавлением необходимых реагентов. Во-вторых, когда комбинаторная библиотека очень велика (и реагенты примерно одинаково реакционоспособны), несколько реакций по созданию комбинаторной библиотеки можно провести в одной ячейке плашки скринера. Тогда биологическому испытанию подвергается вся смесь возможных соединений. Получение положительного сигнала говорит о том, что одно из соединений является активным и далее можно соединения комбинаторной библиотеки получать отдельно.

Первоначально задачей комбинаторной химии являлось создание диверсифицированного набора химических соединений в надежде, что это позволит найти больше хитов в биологических испытаниях. Однако оказалось, что такое тотальное тестирование соединений

комбинаторной библиотеки менее эффективно, чем тестирование хорошо известных и давно полученных соединений. Одна из причин возникновения этой проблемы заключается в том, что соединения комбинаторных библиотек являются в основном более крупными, чем давно известные соединения. Кроме того, соединения больших размеров имеют тенденцию к проявлению большего числа побочных эффектов, содержат нежелательные функциональные группы и проявляют множество нежелательных свойств – плохую растворимость в воде и ДМСО (классических растворителях для скрининга). Это привело к пониманию того, что необходим дизайн комбинаторных библиотек, а именно: для увеличения эффективности скрининга библиотеки должны быть созданы таким образом, что полученные соединения были «лекарствоподобными», а в идеале – показывали активность в моделях, связывающих структуру и свойства. Комбинаторные библиотеки могут быть также созданы под заданную биологическую мишень так, чтобы полученные соединения обладали, по результатам проведения докинга, высоким сродством к биологической мишени, или обладали заданным расположением центров связывания с белком (фармакофором). Кроме того, комбинаторная библиотека должна обеспечивать оптимальный баланс разнообразия и сходства. Рациональный дизайн комбинаторных библиотек является одной из основных задач применения хемоинформатики в комбинаторной химии и биологическом скрининге.

Далее в этом разделе мы коснемся вопросов, касающихся применения хемоинформатики для решения задач комбинаторной химии и создания библиотек соединений для экспериментального и виртуального скрининга.

3.5.1. Перечисление соединений библиотеки

Согласно определению, приведенном на сайте журнала Nature, «*Комбинаторные библиотеки* представляют собой наборы химических соединений, малых молекул или макромолекул, таких как белки, синтезированные комбинаторной химией, в которой множество различных комбинаций родственных химических веществ взаимодействуют друг с другом в сходных химических реакциях»[1]. Если соединения из комбинаторных библиотек были синтезированы в реальном химическом эксперименте, то говорят об

[1] https://www.nature.com/subjects/combinatorial-libraries

экспериментальных комбинаторных библиотеках, а если они были лишь сгенерированы на компьютере, то говорят о *виртуальных комбинаторных библиотеках*.

Создание виртуальной комбинаторной библиотеки соединений и работа с экспериментальными комбинаторными библиотеками требуют использования процедуры *перечисления соединений библиотеки* (англ. library enumeration) – процесса автоматической генерации описывающих химические соединения молекулярных графов.

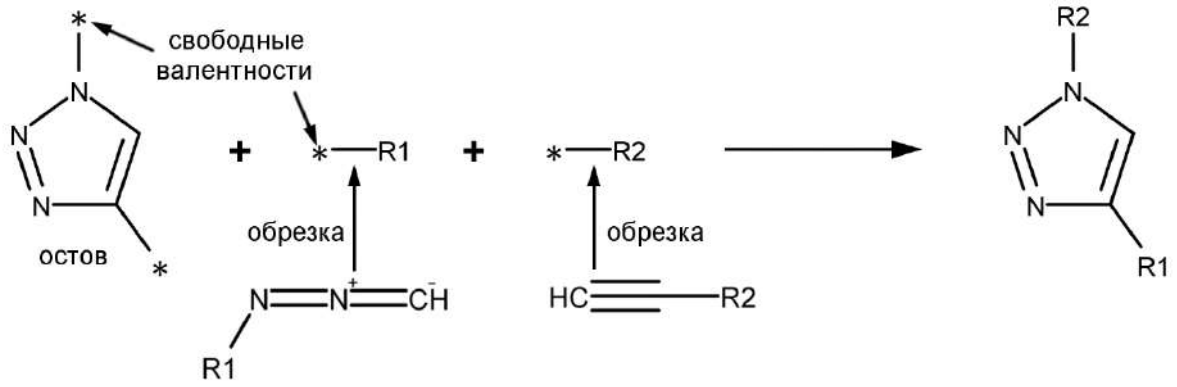


Рис. 52. Иллюстрация к подходу маркирования фрагментов для перечисления соединений библиотеки.

Существует три основных подхода к перечислению соединений комбинаторных библиотек. В первом из них, называемым *маркированием фрагментов* (англ. *fragment marking*), соединения библиотеки считаются состоящими из одного и того же остова (англ. *core template*) и одной или нескольких R-групп. Места их сочленения маркируются путем указания свободных валентностей атомов (Рис. 52). Места сочленений в R-группах можно также указывать не с помощью свободных валентностей, а помечая их символами свободных групп (R1, R2, X1, X2 и т.д.). Такие же символы наносятся и на остов (родительский фрагмент), чтобы указать, куда должны присоединяться R-группы (Рис. 53). Фрагменты для построения соединений строятся из реагентов, удаляя атомы реакционного центра, которые превращаются в родительский фрагмент или не интересующие пользователя продукты реакции. Превращение исходных соединений во фрагменты для перечисления соединений комбинаторной библиотеки называется «*разметкой*» (англ. markup) или «*обрезкой*» (англ. clipping) соединений. Разметку исходных соединений можно проводить в автоматическом режиме. Когда определены остов и R-группы, подход маркирования фрагментов

позволяет вести очень быструю автоматическую генерацию соединений, комбинируя фрагменты, например, с помощью поискового дерева. Такой подход, например, используется бесплатной программой SmiLib [140]. Ограничением этого подхода является то, что не во всех случаях оказывается возможным автоматически определить остов и R-группы.

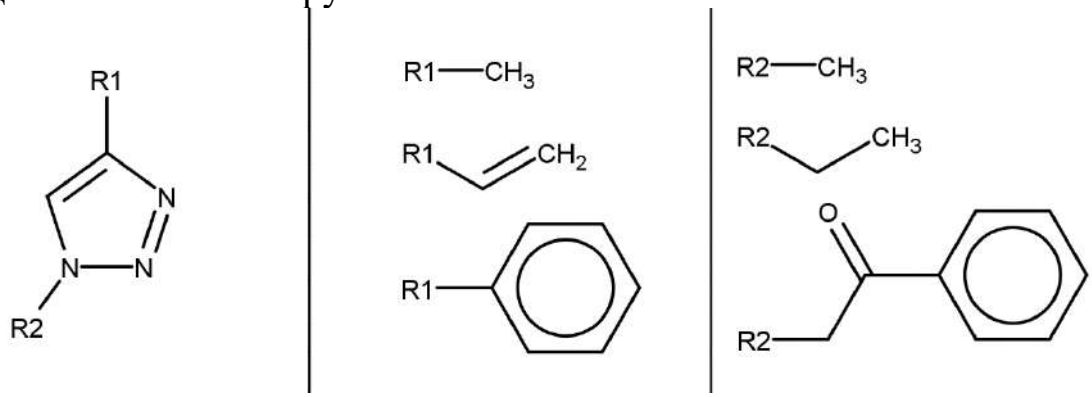


Рис. 53. Разметка фрагментов с использованием R-групп.

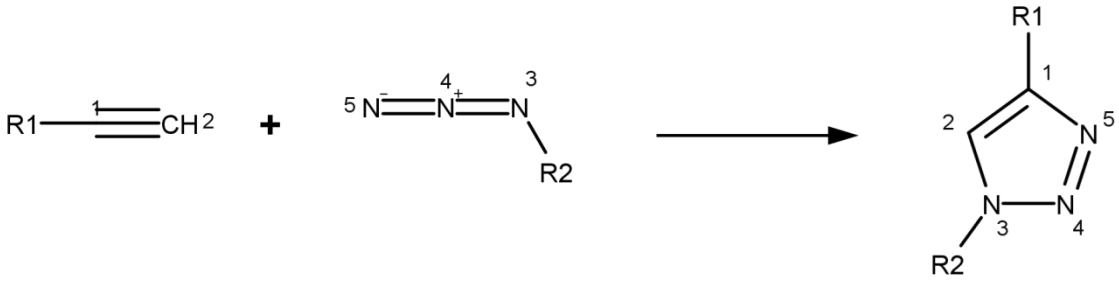


Рис. 54. Иллюстрация к методу реакционных трансформаций для перечисления соединений в библиотеках.

Второй подход к перечислению соединений в комбинаторных (преимущественно экспериментальных) библиотеках называется методом *реакционных трансформаций* (англ. reaction transform). Для этого задается реакционная трансформация – суммарная реакция превращения исходных веществ в продукты, а программа ее использует для генерации всех возможных продуктов. Этот подход требует создания списков реагентов каждого типа, например, для формирования комбинаторной библиотеки тиазолов, исходя из замещенных тиомочевины и α-галогенкетонов (Рис. 54), составляются списки доступных тиомочевин и кетонов. Далее программа перебирает все возможные комбинации соединений, содержащихся в списках, для осуществления генерации конечных соединений. Ключевым моментом этого подхода является необходимость включения атом-атомных отображений в описание трансформаций. Этот метод очень тесно

связан с химическим синтезом, лишен многих недостатков подхода маркирования фрагментов, применим практически ко всем реакциям и понятен химикам. Тем не менее, создание списков реагентов и описание трансформаций для его работы требует определенных навыков и не всегда может быть проведено полностью автоматически.

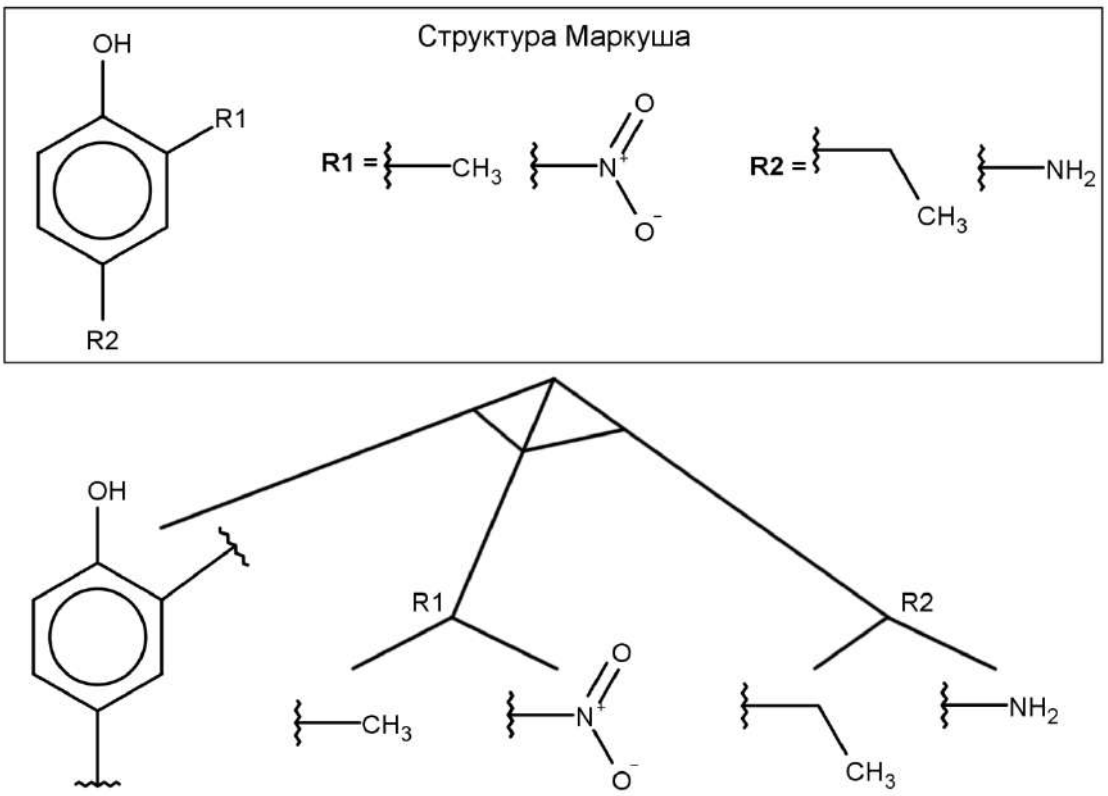


Рис. 55. Описание обобщенной структуры Маркуша 2,4-замещенных фенолов. Каждая вершина-лист представляет собой ее «строительный блок». «Арки» между ветвями означают, что они относятся друг к другу с помощью логической операции И, а при отсутствии «арок» – операции ИЛИ.

Третий подход, предложенный Даунсом и Барнардом [141], предназначен для работы с (виртуальными) комбинаторными библиотеками с использованием представлений соединений в ней с помощью обобщенных структур Маркуша. Для работы с комбинаторными библиотеками в этом случае используются те же подходы, что и при работе с патентными базами данных. Комбинаторная библиотека описывается с помощью обобщенной структуры Маркуша, внутренним представлением которой является так называемое «И/ИЛИ-дерево» - дерево, в котором вершины-листья описывают R-группы и родительскую структуру Маркуша, то есть фрагменты, из которых строятся структуры индивидуальных

химических соединений. Между ветвями «И/ИЛИ-дерева» могут быть определены отношения двух типов (Рис. 55): два ребра могут относиться друг к другу с использованием логической операции И (то есть структурные фрагменты, к которым ведут эти ветви, должны присутствовать в конечной структуре одновременно) или с помощью логической операции ИЛИ (то есть один из фрагментов, к которым ведут эти ветви, должен присутствовать).

Каждый «строительный блок» структуры Маркуша можно описать с помощью битовой строки, составленной на основе специально заданной библиотеки фрагментов. Тогда при осуществлении поиска по сходству, при кластеризации или формировании диверсифицированных наборов для каждой структуры можно очень быстро вычислять фингерпринты структур «на лету» (без составления файла с описаниями фингерпринтов всех соединений из комбинаторной библиотеки), используя логические соотношения между ветвями «И/ИЛИ-дерева». Если «строительные блоки» структуры Маркуша соответствуют ветвям, между которыми установлены И-взаимоотношения, тогда общий фингерпринт химического соединения получается наложением фингепринтов фрагментов с помощью операции И (если бит активирован хотя бы в одном исходном фингепринте – он активируется и в конечном). Если между ветвями определено отношение ИЛИ, то фингерпринт конструируется на основе какого-то одного выбранного фрагмента. Таким образом, метод структур Маркуша позволяет создать очень компактное представление комбинаторной библиотеки, с которым можно легко осуществлять множество операций без необходимости генерировать структуры всех входящих в нее соединений. Такие возможности, например, предоставляет система Torus [142] от Digital Chemistry, а определенную функциональность (создание, перечисление соединений) предоставляют также продукты компании ChemAxon (Marvin Sketch, Markush Viewer, Markush Enumerator).

3.5.2. Дизайн комбинаторных библиотек

Для определения комбинаторной библиотеки необходимо выбрать реагенты для проведения реакций определенного типа и описать, каким образом комбинируются реагенты (обычно используются все возможные сочетания реагентов). Первая возможность заключается в том, что отбираются реагенты для комбинаторного синтеза. Тогда последующему скринингу подвергаются все продукты, получаемые комбинированием этих реагентов. Вторая возможность заключается в

отборе продуктов реакции для экспериментального скрининга с последующим выбором реагентов, ведущих к данному продукту комбинаторного синтеза.

3.5.2.1. Дизайн комбинаторных библиотек, основанный на реагентах

При дизайне комбинаторных библиотек, основанном на реагентах, отбор реагентов ведется без рассмотрения свойств продуктов реакций.

Рассмотрим для примера двухкомпонентную комбинаторную библиотеку, для синтеза которой использовано по 100 реагентов для каждой компоненты, что в сумме дает 10 000 разных продуктов. Допустим, при формировании комбинаторной библиотеки из них нужно отобрать 100 максимально разнообразных соединений. Это означает, что необходимо отобрать по 10 реагентов для каждой компоненты. Вариантов отобрать n реагентов из общего набора, состоящего из N соединений, очень много:

$$\frac{N!}{n!\,(N-n)!}$$

В частности, для рассматриваемой задачи n=10, N=100, и поэтому существует 10^{13} возможных вариантов выбора реагентов для каждого компонента, что дает 10^{26} вариантов для двухкомпонентной комбинаторной библиотеки. Поскольку работать с таким гигантским числом комбинаций невозможно, используются описанные выше приближенные способы отбора реагентов: кластеризация, оптимизация (с помощью генетического алгоритма и искусственного отжига), отбор по несходству.

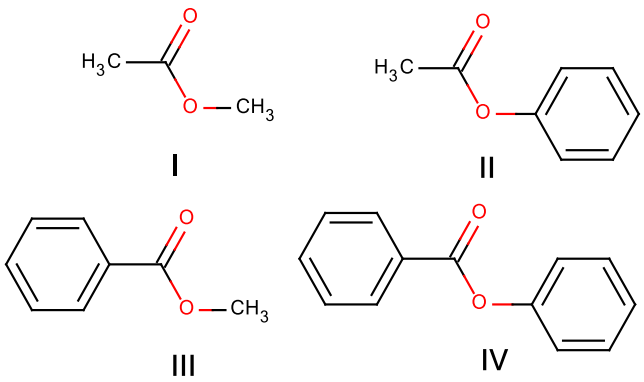


Рис. 56. Возможные продукты, полученные из двух кислот и двух спиртов

В основе рассматриваемого метода лежит предположение о том, что из сильно различающихся реагентов получаются сильно

различающиеся продукты и наоборот. На практике это, однако, не всегда так. Рассмотрим для примера уксусную и бензойную кислоту в качестве одного реагента, а метиловый спирт и фенол – в качестве другого. Хотя структуры этих реагентов сильно отличаются, из возможных продуктов (приведенных на Рис. 56) два продукта – II и III – структурно очень близки между собой, тогда как продукты I и IV отличаются весьма существенно.

Рассматриваемый подход впервые был применен Мартином с соавторами [143] для выбора разнообразных мономеров для комбинаторного синтеза пептоидов[1]. Для расчета меры различия между химическими соединениями были использованы разнообразные дескрипторы (например, ClogP, топологические индексы, фингерпринты и др.). Диверсифицированный набор реагентов формировался с помощью D-оптимального дизайна[2].

---

[1] Пептоиды – это синтетические олигомерные молекулы, образованные α-аминокислотами, в которых боковые цепи присоединены не к α-атому углерода, а к атому азота аминной группы (N-замещенных олигоглицинов).

[2] D-оптимальный дизайн – одна из технологий экспериментального дизайна. Экспериментальный дизайн – технология планирования эксперимента таким образом, чтобы минимальным числом проведенных наблюдений получить максимальный объем объективной информации о влиянии различных факторов на систему. Суть подхода D-оптимального дизайна заключается в поиске такого набора экспериментов (строк, описывающих факторы), чтобы между описывающими его факторами обнаруживались как можно более слабые линейные зависимости, то есть максимизации детерминанта $|\boldsymbol{X}^T\boldsymbol{X}| \rightarrow max$. По этой причине D-оптимальные планы минимизируют ожидаемую ошибку предсказания зависимой переменной, то есть такие плана будут максимизировать точность прогноза, а значит, информацию (которая определяется как обратная величина ошибки), извлекаемую из интересующей нас экспериментальной области. В случае поиска оптимальной библиотеки соединений под факторами понимаются дескрипторы, под экспериментом – соединения, тогда матрица $\boldsymbol{X}$ есть матрица дескрипторов. Таким образом, D-оптимальный дизайн проводит отбор таких соединений в набор, дескрипторы которых как можно хуже скоррелированы.

3.5.2.2. Дизайн комбинаторных библиотек, основанный на продуктах

При дизайне комбинаторных библиотек, основанном на продуктах, при выборе реагентов для синтеза руководствуются свойствами продуктов комбинаторного синтеза.

В рамках данного подхода после того, как оптимальная (сфокусированная или диверсифицированная) библиотека продуктов отобрана любым из описанных выше методов (кластеризацией, оптимизацией, отбором по несходству), могут быть восстановлены реагенты, ведущие к данному набору продуктов. Этот процесс называют «*отбором вишенок*» (англ. cherry-picking).[1] В вычислительном плане этот метод весьма эффективен и интуитивно прост. Для его реализации можно использовать стандартные (описанные выше в разделе 3.4) методы отбора соединений. Единственные отличие от метода отбора по реагенту заключается в необходимости вести отбор из набора соединений, полученных полным перечислением комбинаторной библиотеки, что может быть вычислительно более затратным. В то же время этот подход, является очень неэффективным в синтетическом плане при использовании установок параллельного синтеза. Действительно, для синтеза двухкомпонентной комбинаторной библиотеки, состоящей из 100 соединений, сформированных с помощью рассмотренного выше подхода, основанного на реагентах, надо провести 100 реакций, комбинируя 10 реагентов одного типа с 10 реагентами другого типа, что с помощью параллельного синтеза может быть проведено за 10 шагов. В то же время, для синтеза комбинаторной библиотеки такого же размера, сформированной «отбором вишенок», требуется провести 100 попарных реакций между 100 реагентами одного типа и 100 реагентами другого типа, что не может быть проведено меньше, чем за 100 шагов. В этом случае при проведении параллельного синтеза может быть получено до 10000 продуктов, из которых 9900 окажутся «лишними».

---

[1] Происхождение этого термина, по-видимому, связана с привычкой у детей «выковыривать» вишенки из вишневого пирога, не съедая при этом оставшуюся часть пирога. Следует иметь в виду, что этот термин также нашел широкое применение в логике для обозначения крайне нежелательной практики, когда для доказательства какого-либо утверждения рассматриваются только подтверждающие его факты при полном игнорировании противоречащих ему фактов.

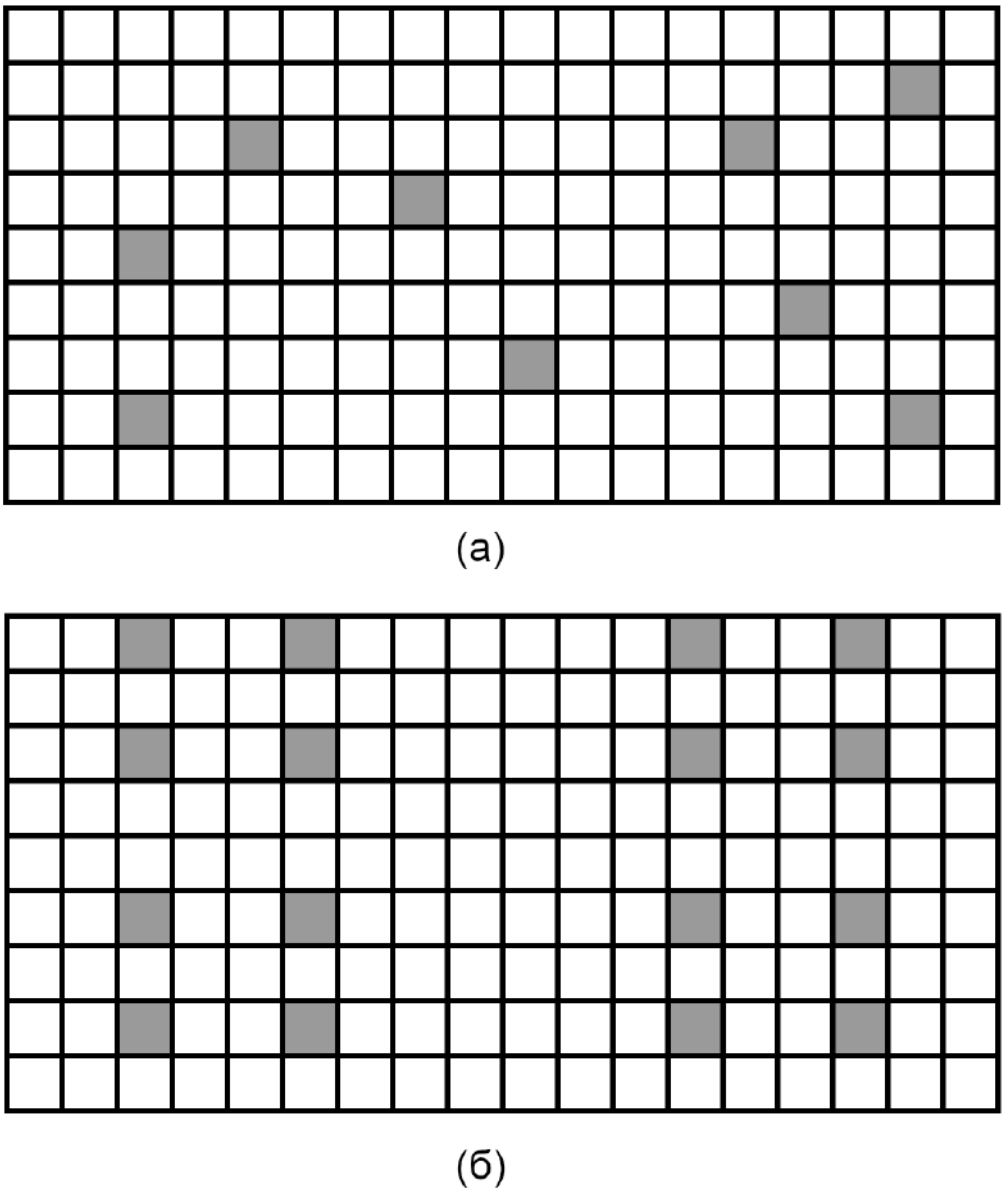


Рис. 57. Примеры матрицы продуктов для (а) разреженного массива (полученного «отбором вишенок») и (б) полного массива. Строки с столбцы соответствуют реагентам двух типов, а заполненные ячейки - отвечающим проведенным с их помощью реакциям.

Сформированные при помощи «отбора вишенок» комбинаторные библиотеки, когда из набора реагентов получаются не все возможные продукты, могут быть представлены с помощью *разреженных массивов* (англ. *sparse array*) – матриц, строки в которых отвечают одному типу реагентов, столбцы – другому, а заполненные ячейки – синтезированным с их помощью продукту (Рис. 57а). Существенно повысить синтетическую эффективность можно, вводя комбинаторное ограничение, предписывающее получать максимально разнообразные

наборы продуктов, вводя в реакцию реагенты одного типа в комбинации со всеми реагентами другого типа. Набор продуктов, которые получаются при проведении реакций между всеми возможными комбинациями реагентов, могут быть представлены *полными массивами* (англ. *full array*) – матрицами, не содержащими пустых ячеек, Рис. 57б. Под дизайном комбинаторных библиотек, основанным на продуктах, обычно понимается именно подход, создающий полный комбинаторный массив. Альтернативный подход обычно называют просто «отбором вишенок», несмотря на то, что он тоже основан на продукте. Мы далее так же будем следовать этой традиции.

Несмотря на существенную вычислительную сложность, основанный на продуктах с использованием полных массивов дизайн комбинаторных библиотек может быть более эффективным методом для оптимизации свойств библиотеки соединений (например, их диверсифицированности по структуре или свойствам). Так, было показано, что отбор по продуктам позволяет создать более диверсифицированные библиотеки, чем отбор по реагентам [144, 145]. Более того, было показано, что основанный на продуктах метод отбора более пригоден для создания сфокусированных библиотек.

Учитывая вычислительную сложность отбора по продуктам, необходимо предельно уменьшить число рассматриваемых соединений. Поэтому обычно в той или иной степени используется трехстадийный алгоритм проведения данного типа отбора. На первой стадии проводится идентификация необходимых реагентов путем отсечения тех из них, которые не удовлетворяют определенным свойствам (например, недоступны для заказа, не имеются в наличии на складе, содержат слишком большие функциональные группы или не совместимы с данной реакционной трансформацией или свойством молекулы, не удовлетворяют «правилу трех»). Далее проводится перечисление всех возможных продуктов (теоретическое перечисление элементов комбинаторной библиотеки). На втором этапе соединения виртуальной библиотеки отсеиваются с помощью различных фильтров, а с использованием моделей «структура-свойство» определяется, какие из них являются более перспективными, и в соответствии с этим проводится ранжирование. На третьем этапе с использованием проведенного ранжирования и/или специальных критериев, таких как необходимая степень разнообразия или сходства соединений, или даже цена соединений, осуществляется отбор соединений в формируемую библиотеку для синтеза. На этом этапе реагенты часто отбираются с использованием метода

искусственного отжига [95, 127, 130, 131, 145, 146] или генетического алгоритма [95, 147].

Рассмотрим в качестве примера метод SELECT [147], базирующийся на генетическом алгоритме, в котором «хромосома» («особь») кодирует набор используемых реагентов. Для этого она разбита на R частей в соответствии с числом реагентов, вступающих в реакцию. В этом случае *i*-ая часть содержит n_i чисел (их можно считать «генами» внутри «хромосомы»), идентифицирующих возможные реагенты *i*-ого типа, которые предполагается использовать для формирования комбинаторной библиотеки. Например, если реакция вовлекает 2 реагента, и требуется отобрать 5 реагентов одного типа и 6 – другого, то «хромосома» состоит из 11 чисел-«генов», идентифицирующих эти реагенты. Функция соответствия (мера «жизнеспособности особи») зависит от разнообразия продуктов, получаемых комбинированием реагентов, и значений определенных свойств химических соединений. Такой подход позволяет получать наборы реагентов для создания максимально разнообразных соединений в комбинаторной библиотеке, которые при этом отвечают определенным физико-химическим ограничениям. Небольшая модификация этих методов (изменение вида функции соответствия, например, путем включения оценки сходства с заданными соединениями-шаблонами, обладающими определенным типом активности) позволяет получать сфокусированные библиотеки.

В программе DirectedDiversity, упомянутой выше в разделе 3.4.4.1, для отбора библиотеки используется метод искусственного отжига [132]. Помимо отбора отдельных соединений (по сути – «отбора вишенок»), эта программа может работать с комбинаторными библиотеками и отбирать полный массив соединений комбинаторной библиотеки, а также отбирать отдельные плашки. В последнем случае предполагается, что библиотека соединений может предоставляться в виде плашек с уже помещенными туда соединениями. Тогда для скрининга необходимо отобрать некоторое число таких плашек, которые обеспечат максимальное разнообразие соединений или, наоборот, максимальную сфокусированность по отношению к некоторой заданному хиту. Работа с полным массивом соединений отличается мало от работы с плашками: перечисляются все соединения, которые им соответствуют, и для них рассчитывается оптимизируемая функция соответствия. В программе DirectedDiversity для этого используется сложная функция, вклад в которую вносят функции, оценивающие разнообразие соединений, сходство с заданным набором соединений, соответствие заданным ограничениям

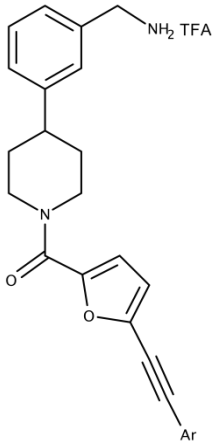
(например, правилу Липинского) и другое. Далее отобранный набор соединений или плашек «мутируется», то есть некоторое число объектов заменяются другими из исходного набора. Если данное замещение приводит к улучшению значений функции соответствия, данное изменение принимается, в противном случае используются критерии принятия (например, критерий Метрополиса). Со временем вероятность принятия трансформаций наборов, приводящих к ухудшению значений функции соответствия, уменьшается (с помощью уменьшения «температуры») и решение сходится.

Заметим, что изложенные выше методы оптимизируют всю библиотеку целиком. Каждое из рассматриваемых решений представляет собой отобранную библиотеку, поэтому такие методы называют *основанными на библиотеке*. Для создания сфокусированных библиотек с использованием генетического алгоритма (или любого другого алгоритма, одновременно рассматривающего семейство решений) могут также использоваться методы отбора, основанные *на индивидуальных соединениях* [148]. Для них каждое рассматриваемое решение представляет собой конкретное химическое соединение. Так, в методе Шеридана-Кирсли [149], осуществляющем дизайн комбинаторных библиотек, каждая «хромосома» описывает структуру входящего в библиотеку соединения как комбинацию фрагментов (напомним, что в методе SELECT «хромосома» представляет всю отобранную библиотеку). Функция соответствия зависит от сходства структуры с заданной. Таким образом, процесс оптимизации подбирает набор соединений, наиболее сходных с заданным. После завершения процедуры генетической оптимизации программой определяются фрагменты, которые наиболее часто встречающиеся в полученном в результате оптимизации наборе соединений. Эти фрагменты используются для определения наилучшей сфокусированной комбинаторной библиотеки. Преимуществом этого метода является относительная экономичность расчетов, поскольку функция соответствия вычисляется относительно быстро (особенно если она основана на 2D свойствах). Кроме того, этот метод позволяет провести дизайн сфокусированной комбинаторной библиотеки после исследования относительно малой части поискового пространства.

3.5.2.3. Дизайн комбинаторных библиотек, основанный на структуре биомишени

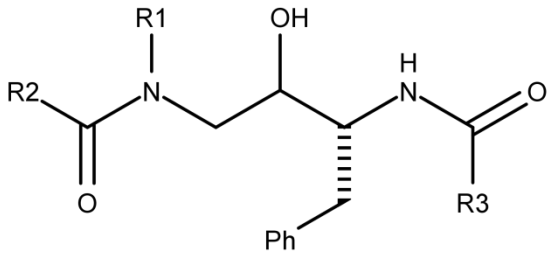
В том случае, когда известна структура биомишени, становится возможным осуществлять дизайн комбинаторных библиотек с учетом этой информации. Зачастую для этого используют технологию виртуального скрининга: перечисляются все соединения некоторой комбинаторной библиотеки, рассчитываются дескрипторы, при необходимости соединения фильтруются, в том числе это может быть сделано с использованием подготовленных моделей «структура-свойство», и с использованием докинга или фармакофорного поиска отбираются наиболее подходящие соединения. Структура реагентов восстанавливается и формируется разреженный комбинаторный массив для биологических испытаний.

Примером такого подхода может служить создание противоаллергенных препаратов на основе ингибирования медиатора воспаления β-триптазы [150]. В этой работе, исходя из структуры лигандов и их расположения в сайте связывания белка, было решено искать оптимальный препарат в виде следующей обобщенной структуры:



Исходя из предполагаемой стратегии синтеза этих замещенных, был проведен поиск возможных реагентов в коммерческих базах данных, и на их основе сформирована комбинаторная библиотека. Был проведен докинг соединений в активный сайт β-триптазы, из них 59 соединений смогли быть размещены в полости фермента и 25 соединений, показавших наибольшее сродство к белку, были переданы на биологические испытания. Этот подход показал очень высокую эффективность: 17 соединений показали субмикромолярную константу ингибирования, 2 соединения – подходящий фармакокинетический профиль и одно соединение было отправлено на клинические испытания.

В классической работе Кирка с соавт. [151] по созданию комбинаторных библиотек на основании знания структуры белка катепсина Д, имеющего отношение к болезни Альцгеймера и некоторым формам рака, и его комплексов с ингибиторами, был предложен остов (скаффолд) **1**.



1

Для формирования комбинаторной библиотеки были выбраны подходящие для синтеза и доступные на рынке амины (~700) и карбоксилирующие реагенты (~1900).

Для уменьшения сложности задачи проведения докинга было предположено, что остов должен располагаться в сайте связывания белка точно таким же образом, как соответствующий фрагмент природного ингибитора пепстатина в комплексе катепсина, и в процессе докинга не изменяет своего положения. Часть остова, включающая заместители R_1 и R_2, не участвует в прочном связывании с белком и поэтому, как было предположено, может располагаться в комплексе с белком произвольным образом. Большая часть конформаций при этом имеют близкую энергию, и при этом было показано, что конформации остова образуют 4 кластера. Таким образом, существует 4 варианта расположения этого остова с заместителями в полости фермента, что существенно сокращает сложность задачи, поскольку, несмотря на наличие множества степеней свободы остова, выбор идет только между четырьмя вариантами. Тогда задача докинга сводится к последовательному добавлению заместителей, помещенных в точках вариации родительского фрагмента, перебору конформаций только варьируемых фрагментов и расчету функций скоринга (некоторого подобия свободной энергии связывания[1]). Этот подход используют в дизайне

[1] На самом деле несмотря на то, что поначалу функции скоринга представляли некоторый способ оценки (свободной) энергии связывания, в настоящее время их можно рассматривать как показатель, отражающий,

комбинаторных библиотек лигандов на основании структуры биомишени [152]. В этой работе добавление заместителей и докинг проводилось с помощью программы CombiBuild, варианта программы BUILDER [153], а использованная функция скоринга была основана на силовом поле AMBER. При этом проводился полный перебор конформеров заместителей, а чтобы облегчить задачу конформационного поиска, были удалены фрагменты молекул, имеющие более 4 свободно вращающихся связей. В результате этого 50 соединений с наилучшими значениями функции скоринга были отобраны для дальнейшей работы.

Из отобранных структур работе Кирком с соавт. [151] были восстановлены структуры необходимых для их синтеза реагентов, из которых слишком дорогие были удалены. В итоге осталось 34, 35 и 41 реагентов, соответствующих заместителям R_1, R_2 и R_3. Полный массив комбинаторной библиотеки тогда содержал бы почти 50 000 продуктов. Авторам необходимо было отобрать только 1000 вариантов, что соответствует комбинаторной библиотеке, сформированной путем комбинирования 10 реагентов каждого типа. Поэтому авторы поступили следующим образом: структуры реагентов, соответствующих каждому из трех заместителей, были подвергнуты иерархическому кластерному анализу (использовались молекулярные отпечатки Daylight и индекс Танимото) и выбран уровень сходства, на котором было выделено по 10 кластеров. Из каждого кластера было выбрано по одному соединению. Полный массив, соответствующий данной комбинаторной библиотеке, был синтезирован согласно предложенному синтетическому плану и подвергнут биологическому скринингу. В ходе скрининга было обнаружено 67 соединений с субмикромолярной активностью (то есть $IC_{50}<10^{-6}$ моль/л[1], 7 соединений, для которых $IC_{50}<10^{-7}$ моль/л).

Для сравнения авторы создали диверсифицированную библиотеку соединений без знания структуры биомишени. Для перечисления (генерации) соединений использовался исходный набор реагентов из ~900 аминов и ~1900 карбоксилирующих агентов. Структуры реагентов из каждого набора были подвергнуты кластерному анализу с использованием метода Ярвиса-Патрика и

насколько лиганд хорошо «подходит» данному белку. Многие современные функции скоринга не несут смысла энергии.

[1] IC_{50} – характеристика ингибирующей активности соединения, представляющая собой концентрацию ингибитора, при которой активность фермента составляет 50% от исходной.

метрики Танимото на молекулярных отпечатках Daylight. Соединения, расположенные наиболее близко к центроиду каждого кластера, были выбраны в качестве его представителей. Всего было выбрано 47, 23 и 35 реагентов, соответствующих заместителям R_1, R_2 и R_3. Из них вручную было отобрано по 10 представителей так, чтобы обеспечить приемлемую стоимость и достаточное структурное разнообразие. После синтеза соединений библиотеки и проведения биологических испытаний оказалось, что 26 соединений показали субмикромолярную активность и только 1 имело $IC_{50}<10^{-7}$ моль/л.

Таким образом, было продемонстрировано, что дизайн комбинаторных библиотек с использованием информации о структуре биомишени значительно более эффективен, чем без использования этой информации. Эта работа показала, какое значение имеет интегрирование комбинаторной химии с методами дизайна, основанных на структуре биомишени [152]. К настоящему времени опубликовано и множество других работ, основанных на этом подходе. В качестве примера можно привести дизайн ингибиторов плазмепсина II [154], тромбина [155], аспарагиновой пептидазы [156], энтеротоксина кишечной палочки [157], блокаторов калиевых каналов Т-клеток [158].

К. Хьюэ [159], обобщив предложенные ранее подходы к дизайну комбинаторных библиотек с использованием информации о структуре биомишени, предложил гибкий метод, включающий: выбор исходной синтетической концепции, виртуальный скрининг соединений комбинаторной библиотеки, улучшение свойств соединений с помощью оптимизации фармакокинетических и фармакодинамических характеристик, формирование конечной библиотеки с помощью многоцелевой оптимизации и окончательный отбор соединений (Рис. 58).

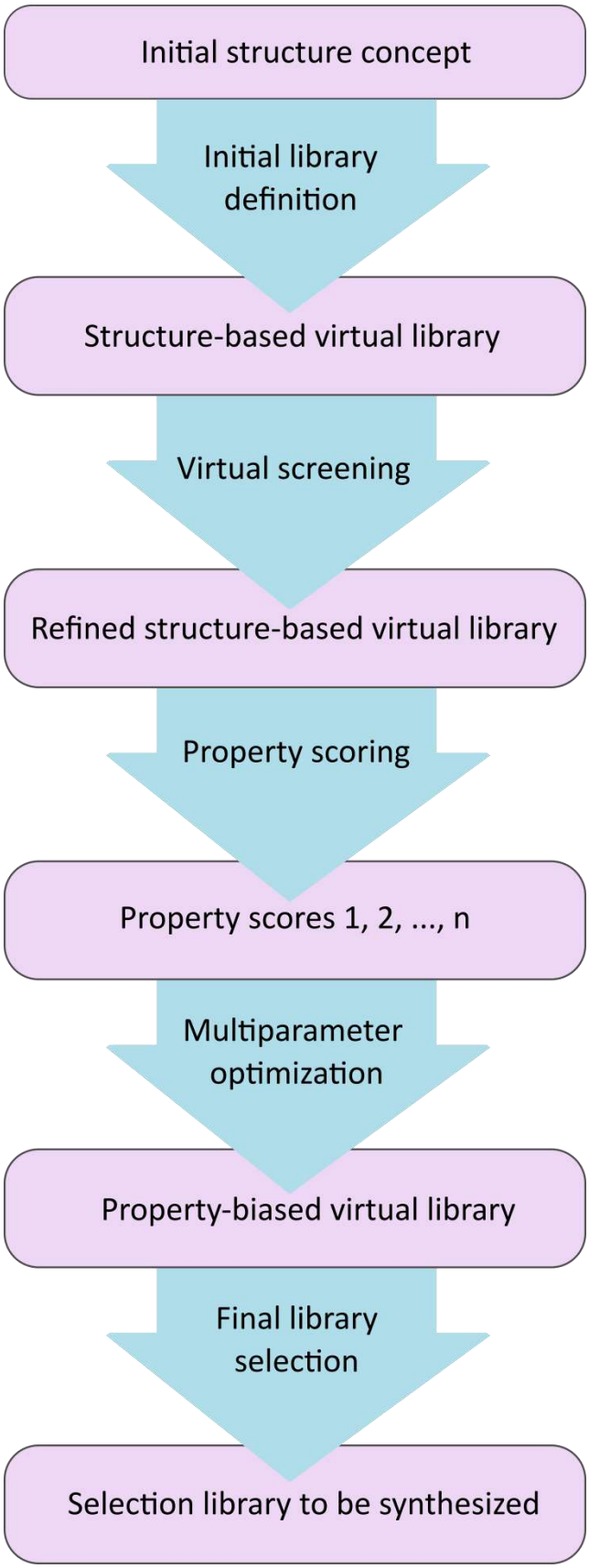


Рис. 58. Подход к компьютерному дизайну комбинаторных библиотек с использованием информации о структуре биомишени.

4. ФАРМАКОФОРНЫЙ АНАЛИЗ

Молекулярная структура органического лиганда может быть представлена в виде набора *фармакофорных признаков* (англ. pharmacophore features), ответственных за различные взаимодействия с биомишенью: доноры и акцепторы водородных связей, центры положительного и отрицательного заряда, гидрофобные и ароматические группировки. Фармакофорные признаки также часто называют *фармакофорными центрами*. Набор фармакофорных признаков, присутствующих в молекуле, формирует ее *фармакофорное представление*. Под *фармакофором* (англ. pharmacophore) понимается такой набор расположенных определенным образом фармакофорных признаков, который является общим для молекул, обладающих биологической активностью данного типа. Если такое расположение рассматривается в трехмерном пространстве, то говорят о *трехмерных (3D) фармакофорах*, а если внутри молекулярных графов – о двухмерных (2D) или *топологических фармакофорах*. Взаимное расположение фармакофорных признаков обычно задается путем указания допустимых интервалов расстояний (геометрических для 3D-фармакофоров и топологических для 2D-фармакофоров) между ними, а абсолютное – с помощью радиуса и декартовых координат центра сферы, внутри которой относящимся к этому признаку атомам «разрешено» или «запрещено» находиться. Для обозначения ориентации признаков в 3D-фармакофорах могут быть указаны дополнительные геометрические характеристики, такие как направление образования водородных связей, ориентация плоскости ароматического кольца и другие.

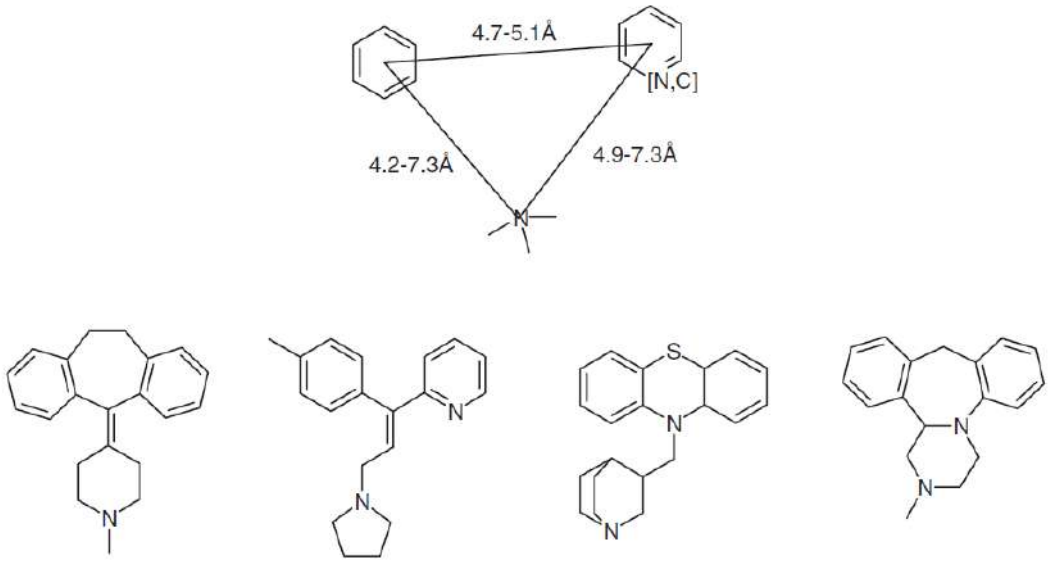


Рис. 59. Трехмерный фармакофор для поиска ингибиторов H1 гистаминных рецепторов, и удовлетворяющие ему молекулы. Рисунок из статьи [160] публикуется с разрешения издательства. Copyright (1995) American Chemical Society.

Работа с трехмерными фармакофорами обычно включает две основные стадии: определение (построение) фармакофора (фармакофорное отображение, англ. *pharmacophore mapping*, или построение фармакофорной модели, англ. *pharmacophore modeling*) и фармакофорный поиск. Подробно фармакофорный поиск изложен в Пособии 2. Термин «фармакофор» в современном его понимании был введен в употребление Монти Киром (Monty Kier) в конце 60-ых годов XX века[1], и им же в 1967 г. был построен первый фармакофор для мускариновой активности, хотя корни этого понятия часто связывают с именем жившего еще в конце XIX и начале XX века основоположника химиотерапии Пауля Эрлиха (Paul Ehrlich). Поскольку построение фармакофора является результатом моделирования, в последнее время, наряду с термином «фармакофор», употребляют имеющий тот же самый смысл термин «*фармакофорная модель*» (англ. pharmacophore model).

Существует два основных способа определения трехмерных фармакофоров (построения фармакофорных моделей): исходя из структур комплексов белок-лиганд (англ. structure-based) или на основе обобщения пространственных структур химических соединений, обладающих определенным типом биологической активности (англ.

[1] См. J. H. Van Drie, Monty Kier and the Origin of the Pharmacophore Concept, *Internet Electron. J. Mol. Des.* **2007**, *6*, 271–279, http://www.biochempress.com.

ligand-based). Далее мы рассмотрим, каким образом ведется построение трехмерных фармакофоров. Топологические фармакофоры могут также создаваться, исходя из трехмерных, для упрощения и ускорения поиска в базах данных на одном из этапов виртуального скрининга. Для этого геометрические расстояния заменяются топологическими.

4.1. ОПРЕДЕЛЕНИЕ ФАРМАКОФОРОВ ИСХОДЯ ИЗ СТРУКТУР КОМПЛЕКСОВ БЕЛОК-ЛИГАНД

Трехмерный фармакофор может быть определен исходя из экспериментально изученной трехмерной структуры комплекса биологической макромолекулы с органическим лигандом. Из анализа структур таких комплексов можно определить, какие взаимодействия обусловливают связывание малых молекул с биологической макромолекулой. Необходимые для этого экспериментальные (данные рентгеноструктурного анализа либо ЯМР) структуры белков и их комплексов могут быть, например, взяты из публично доступного Банка данных белков (*Protein Data Bank, PDB*). Фармакофорное моделирование с использованием данных о структуре белков является комплементарной докингу и при этом менее вычислительно сложной процедурой *виртуального скрининга, основанного на структуре биомолекулы*[1] (англ. structure-based virtual screening).

Существует несколько способов определения фармакофоров на основе структур комплексов белок-лиганд. Желательно, чтобы фармакофоры строились, исходя из структур нескольких комплексов одного и того же белка с различными лигандами. В противном случае построенный фармакофор может оказаться слишком специфичным в скрининге, т.е. с его помощью будут отбираться только соединения, очень похожие на лиганд в комплексе, тогда как молекулы даже с незначительно отличающимся способом связывания с биологической мишенью отбираться не будут.

Простейшим способом определения фармакофоров является анализ сайта связывания, проводимый пользователем «вручную» с использованием программ интерактивной графики. В этом случае

[1] Виртуальный скрининг, основанный на структуре биомолекулы (часто слово «биомолекула» опускается) – это проводимая с помощью компьютера процедура отбора (скринирования) химических соединений на основе вычислений, требующих знание трехмерной структуры сайта связывания биологической мишени.

можно указать необходимые фармакофорные центры в соответствии с тем, какие присутствуют взаимодействия между молекулой низкомолекулярного лиганда и белком. Это может быть также сделано автоматически с использованием определенных правил, которые идентифицируют и классифицируют группы атомов лиганда, взаимодействующие с белковой макромолекулой. При наличии нескольких экспериментально изученных структур комплексов, фармакофоры, построенные на основании структур нескольких комплексов, могут быть объединены в один общий фармакофор. Для этого проводится *выравнивание* (англ. alignment) фармакофоров, то есть находится способ максимально эффективно совместить их в пространстве. Далее, в конечный фармакофор могут отбираться либо общие для всех фармакофоров признаки, либо все фармакофорные признаки объединяются в общий фармакофор. Также общий фармакофор может быть найден выравниванием сайтов связывания белков путем совмещения α-углеродных атомов основной цепи с последующим построением единого фармакофора. Такие алгоритмы реализованы, например, в программе LigandScout [161, 162].

Существуют также другие способы определять фармакофоры исходя из структур комплексов белков с органическими молекулами. Например, Ортузо создал фармакофорную модель, основанную на расчете молекулярных полей в программе GRID [163]. В рамках этого подхода при помощи программы GRID строятся молекулярные поля лиганда и сайта связывания белка. Комбинируя эти поля, можно идентифицировать области пространства, в которых реализуются энергетически выгодные лиганд-белковые взаимодействия различных типов. Большое число возможных проб в программе GRID позволяет сформировать достаточно селективный фармакофор для виртуального скрининга. Используется также подход, при котором различные молекулярные пробы размещаются в различных положениях внутри сайта связывания биологической макромолекулы, на основе чего создается описание молекулярных полей, из которых далее выводится фармакофор [164].

4.2. ОПРЕДЕЛЕНИЕ ФАРМАКОФОРОВ ИСХОДЯ ИЗ СТРУКТУР КОМПЛЕКСОВ БЕЛОК-ЛИГАНД

Процесс определения фармакофоров путем сравнения фармакофорных представлений отдельных молекул называется *фармакофорным отображением* (англ. *pharmacophore mapping*). Фармакофорное отображение позволяет сформулировать запрос,

который может быть использован для отбора из баз данных структур молекул, активных по отношению к заданной биологической мишени (лигандов). Данный подход не использует информацию о структуре биологической мишени или ее активного центра.

При проведении фармакофорного отображения необходимо принимать во внимание три ключевых момента. Первый касается конформационной гибкости молекул. Необходимо принимать во внимание тот факт, что, как правило, трехмерный фармакофор содержится только в части из общего числа конформаций молекулы. Вторая проблема заключается в том, что существует много возможных комбинаций фармакофорных групп молекулы, причем возможно соответствие одной атомной группировки нескольким фармакофорным центрам. Третий момент заключается в том, что один и тот же сайт связывания может взаимодействовать с разными лигандами посредством различных взаимодействий. Дело в том, что сайт связывания биологической мишени может содержать большое количество функциональных групп, и при этом одна из взаимодействующих с ним молекул лиганда может связываться только с частью из них, а другая молекула – с другой комбинацией групп. В связи с этим для построения фармакофора необходимо отбирать структурно однородные молекулы. Заметим, что даже очень похожие молекулы могут связываться совершенно различным образом. На Рис. 60 показан хрестоматийный случай, когда относительно малые изменения в структуре ингибитора дигидрофолата – метатрексана – приводят к тому, что наложение молекул с целью получения общего фармакофора должно происходить совершенно различным и на первый взгляд неочевидным образом.

Задачей фармакофорного отображения является поиск фармакофора, который наиболее хорошо представляет заданный набор активных молекул. Для этого в фармакофор должно отбираться максимальное число фармакофорных признаков, общих для активных молекул. Разработано много методов осуществления фармакофорного отображения: ограниченный систематический поиск, метод на основе рассмотрения способов пересечения графов при помощи поиска клик графа совместимости, метод максимального сходства, методы стохастического поиска. Ниже мы рассмотрим некоторые из них.

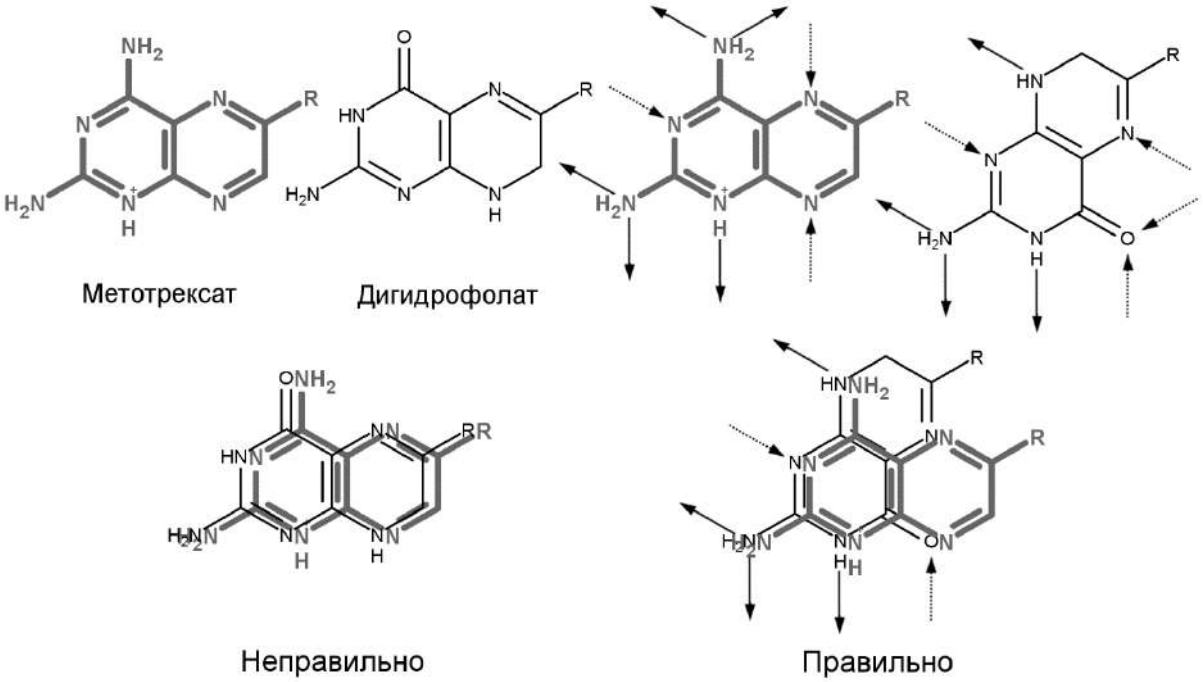


Рис. 60. Кажущееся очевидным наложение метотрексата и дигидрофолата (слева) не обеспечивает оптимальное перекрывание фармакофорных представлений этих молекул. С учетом возможности осуществления одинакового типа связывания с белком они должны перекрываться другим образом (справа).

4.2.1. Фармакофорное отображение с использованием ограниченного систематического поиска

Фармакофорное отображение с использованием ограниченного систематического поиска [165] основывается на алгоритмах систематического исследования конформационного пространства. Как уже упоминалось в Пособии 1, систематический перебор всех комбинаций двугранных углов в молекуле является NP-сложной проблемой, и время, требуемое на расчет, растет экспоненциально с увеличением числа вращающихся связей. Соответственно и сопоставление фармакофоров во всех конформерах молекулы может стать вычислительно очень сложной задачей. Тем не менее, использование «*обрезки*» (англ. pruning) поискового дерева в соответствии с требованиями, которые накладывает необходимость совмещения фармакофоров в молекуле, позволяет проводить этот процесс весьма быстро и эффективно.

На первом этапе пользователь указывает, какие фармакофорные признаки в каждой из сопоставляемых структур должны быть совмещены в конечном фармакофоре. Таким образом, задача фармакофорного отображения сводится к тому, что нужно найти такие диапазоны расстояний между фармакофорными признаками, чтобы полученный фармакофор соответствовал части конформеров (хотя бы одному) каждой из активных молекул обучающей выборки. В начале процедуры фармакофорного отображения отбирается наиболее конформационно жесткая молекула, проводится систематический перебор ее конформаций в пределах заданного энергетического окна и регистрируются расстояния между фармакофорными признаками. В следующей по жесткости молекуле исследование конформационного пространства проводится с использованием полученных для первой молекулы диапазонов расстояний. Таким образом, становится понятным, какие расстояния между фармакофорными признаками реализуются во второй молекуле, а какие не могут реализоваться. Процедура повторяется для всех активных молекул обучающей выборки в порядке роста их конформационной лабильности. Чем больше молекул последовательно включается в рассмотрение, тем более точно находятся «разрешенные» диапазоны расстояний (Рис. 61) и тем более точно будут отбираться молекулы в виртуальном скрининге.

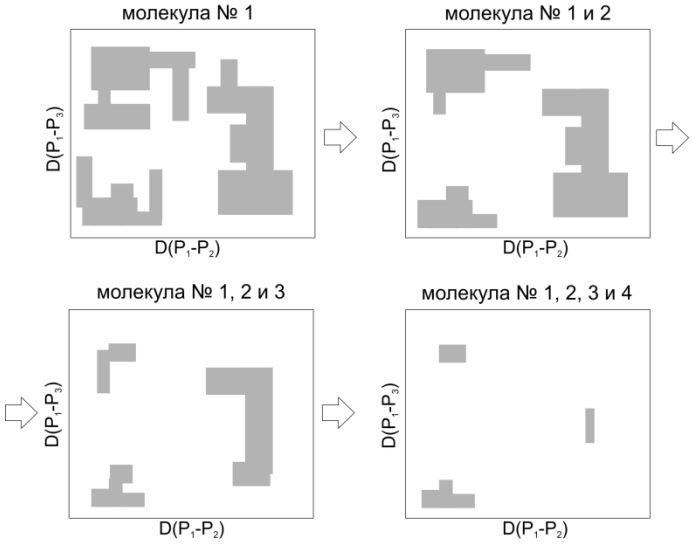


Рис. 61. Уточнение фармакофора с использованием систематического конформационного поиска. Последовательное использование на ряде молекул систематического конформационного поиска с «обрезкой» поискового дерева позволяет уточнить расстояние между фармакофорными центрами P_1, P_2 и P_3 и таким образом найти трехмерный трехточечный фармакофор.

Недостатком данного метода является необходимость указания в структурах молекул фармакофорных признаков на раннем этапе, когда фармакофор еще неизвестен. При наличии достаточно большого количества потенциальных фармакофорных признаков это становится сложно сделать, что приводит к появлению множества возможных фармакофорных моделей. По этой причине данный подход не получил широкого распространения. В ранних работах этот подход назывался *методом активного аналога* (англ. active analog approach) [166].

4.2.2. Фармакофорное отображение с использованием стохастических подходов

Из различных стохастических подходов наиболее широко используется для фармакофорного отображения генетический алгоритм. Он взят за основу во множестве программ: GASP [167-170], GALAHAD [171, 172], MOGA [173-175], GAPE [176] – первые две из которых доступны в составе продуктов компании Tripos.

В отличие от большинства программ, в которых фармакофорное отображение и генерация конформеров являются разделенными процессами, в программе GASP конформационный поиск сопряжен с построением фармакофоров. Весь набор из N молекул кодируется в одной хромосоме, которая состоит из двух основных частей.

В первой части, состоящей из комбинации N битовых строк разной длины, кодируются конформации всех молекул в наборе. Для кодирования каждого двугранного угла в молекуле резервируют 8 бит, что позволяет разбить диапазон его значений (360 градусов) на 28 = 256 корзин. Другая часть хромосомы состоит из комбинации *N-1* строк, содержащих номера фармакофорных признаков данной молекулы, соответствующих таковым в базовой молекуле, Рис. 62. Базовой молекулой считается та, которая содержит наименьшее число фармакофорных признаков. Число элементов в строке соотнесения фармакофорных признаков равно числу признаков в базовой молекуле. Каждый элемент строки равен номеру фармакофорного признака в данной молекуле, который соответствует данному признаку в базовой молекуле. Для генетической оптимизации создается определенное количество таких хромосом, в которых случайным образом распределяются величины углов и соответствий между фармакофорными признаками молекул набора и базовой молекулы. Генетическая оптимизация направлена на улучшение соответствия между фармакофорными признаками различных молекул путем

минимизации суммы квадратов расстояний между ними при наложении молекул в пространстве.

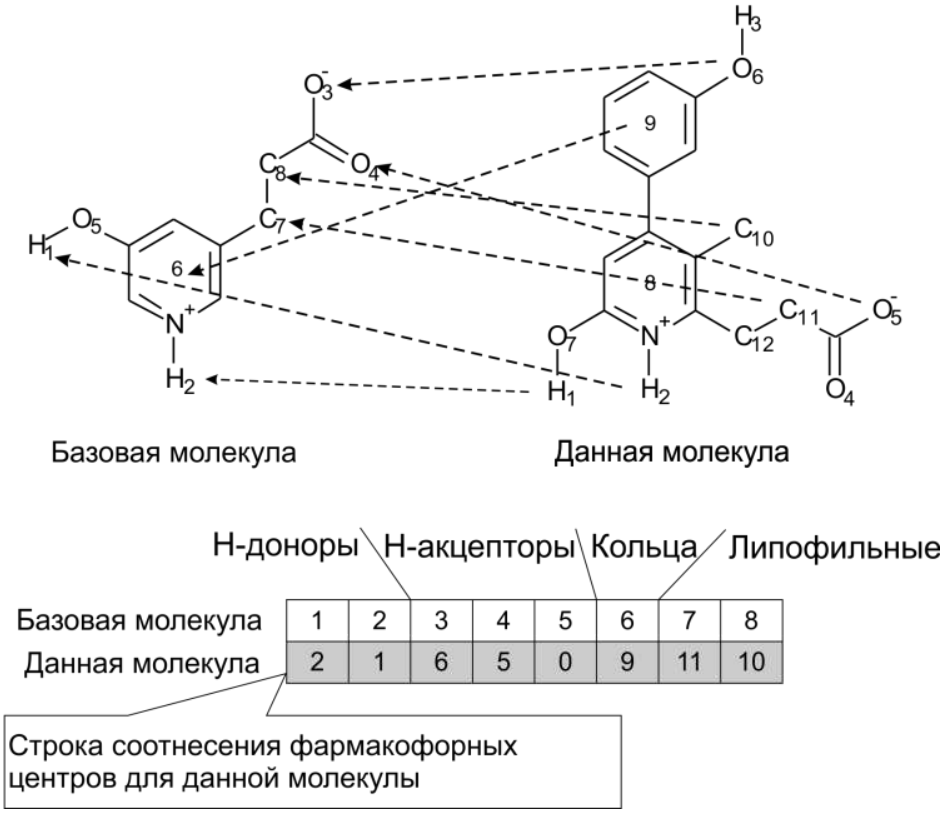


Рис. 62. Создание строки соотнесения фармакофорных центров для данной молекулы в программа GASP. В строке фигурируют номера каждого фармакофорного центра данной молекулы, соотвествующий фармакофорному центру (признаку) базовой.

«Жизнеспособность» хромосом определяется с помощью функции приспособленности, которая в случае GASP состоит из следующих компонентов: (i) сходства в расположении, ориентации и типе H-доноров, H-акцепторов и расположении и ориентации колец, (ii) общего объема наложенных структур, (iii) внутренней стерической энергии конформеров для данной конформации. Далее популяция «особей» с данными хромосомами эволюционирует, чтобы оставить хромосомы, обеспечивающие максимальную жизнеспособность особей (минимальное значение функции приспособленности). В ходе эволюции на хромосомы действует два оператора – кроссинговер и мутация. Для кроссинговера из популяции хромосом отбираются две хромосомы, определяется положение в хромосомах, по которым происходит кроссинговер, по данному месту хромосомы разрезаются и обмениваются получившимися участками. Таким образом, из двух хромосом-родителей получаются две хромосомы детей. Отбор родителей ведется с пользованием алгоритма колеса рулетки, который

гарантирует, что более жизнеспособные родители с большей вероятностью произведут детей. Например, если для трех особей функция приспособленности равна 1, 2 и 3, то на рулетке из 6 (1+2+3) секторов первой особи предоставляется один сектор, второй – 2, а третьей – 3, и вероятность того, что данная особь будет отобрана, равна 1/6, 2/6 и 3/6 соответственно. Для проведения мутации из набора отбирается хромосома, на ней случайно выбирается положение, значение в котором заменяется случайным (отличающимся от существующего). Это означает, что либо заменяется бит с 0 на 1 или наоборот, либо целое число из второй части хромосомы заменяется случайным образом на другое допустимое целое число (то есть не превышающее число фармакофорных признаков в данной молекуле). Если в последнем случае возникает конфликт (то есть данному фармакофорному признаку данной молекулы уже соответствует какой-то признак в базовой молекуле), то процедура повторяется. Таким образом, мутация родителя приводит к одной дочерней хромосоме. Генетический алгоритм, реализованный в программе GASP, использует постоянное возобновление популяции, то есть на каждом шаге генерируется определенное небольшое число потомков, и такое же число плохо приспособленных особей удаляется из популяции. При многопроцессорной реализации используется островная модель: популяция разбивается на отдельные острова, на каждом из которых эволюция идет своим ходом. Время от времени одна особь мигрирует с острова на остров, то есть хромосома из одного набора, отбираемая в соответствии с алгоритмом колеса рулетки, вводится в другой набор. Расчет продолжается до тех пор, пока не пройдет определенное число шагов или в ходе эволюции в течение достаточно большого числа шагов не будет наблюдаться улучшение значений функции приспособленности, превышающее заданный порог.

Лучшая хромосома, полученная в результате эволюции, представляет собой наилучшее фармакофорное наложение молекул обучающей выборки в низкоэнергетических конформациях. Построенная фармакофорная модель используется далее для проведения скрининга – если скринируемая молекула имеет в пределах определенного энергетического окна конформер, который может быть совмещен с фармакофором, то эта молекула отбирается. Данный метод приводит к построению нескольких моделей, которые ранжируются в соответствии с функцией приспособленности. Для осуществления виртуального скрининга можно использовать несколько из построенных фармакофорных моделей.

GALAHAD в целом весьма похож на GASP, но использует большее количество видов фармакофорных признаков, чем GASP. Алгоритм эффективного выравнивания молекул LAMDA используется для вычисления сходства фармакофоров [171]. Процесс расчета разбивается на две стадии. На первой стадии молекулы выравниваются друга относительно друга в заданных трехмерных конформациях (которые либо импортируются, либо генерируются). На этом этапе генетический алгоритм используется для генерирования конформаций лиганда, при которых минимизируется энергия лигандов и максимизируется сходство фармакофорных представлений молекул набора, которые совмещаются в пространстве на втором этапе.

4.2.3. Фармакофорное отображение с использованием поиска клик графа совместимости

Клика (англ. clique) – это подграф в неориентированном графе, представляющий собой *полный граф* (англ. complete graph, то есть все его вершины связаны между собой ребрами), см. раздел 2.3.3. Пособия 2. *Максимальная клика* (англ. maximal clique) – это клика, не содержащаяся внутри какой-либо другой клики (Рис. 63). Любой граф, не являющийся полным, содержит несколько максимальных клик. *Наибольшая клика* (англ. maximum clique) – это максимальная клика, содержащая наибольшее число вершин в графе (Рис. 63). В общем случае в графе может содержаться как одна, так и несколько наибольших клик.[1]

К. Брон и Дж. Кирбош [177] разработали эффективный алгоритм для поиска клик в графе, который активно используется в разных областях хемоинформатики. Например, широкое применение он находит для поиска максимальных общих подструктур в наборе химических соединений.

Поиск максимальных клик для построения общего для набора активных молекул трехмерного фармакофора используется в алгоритме DISCO [178, 179], ставшем основой программы DISCOtech [180], распространяемой компанией Tripos.

---

[1] Следует отметить, что далеко не все авторы строго следует приведенной терминологии, принятой в теории графов. В частности, во многих публикациях наибольшие клики называются максимальными, а максимальные клики – просто кликами.

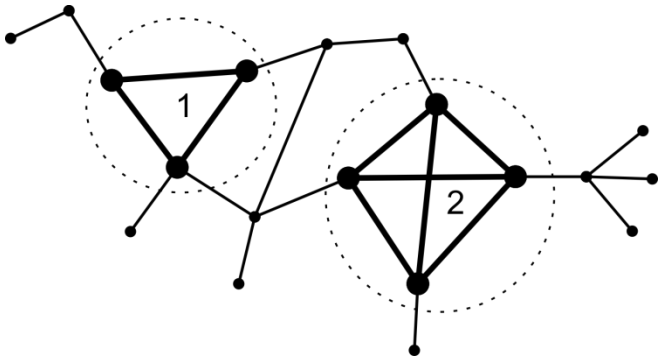


Рис. 63. В указанном графе есть две максимальные клики 1 и 2 с количеством вершин больше двух (клики из двух вершин – то есть просто две связанные вершины – здесь не указаны, поскольку они тривиальны), из которых клика 2 является наибольшей.

В рамках данного подхода на первом этапе ведется генерация заданного числа по возможности разнообразных конформеров для каждой молекулы из набора. Разнообразие конформеров контролируется с использованием вычисления среднеквадратичного отклонения расстояний (RMSD) между всеми парами атомов и, если для какой-нибудь пары конформеров эта величина оказывается меньше порогового значения (в оригинальном алгоритме 0.4Å), то конформер из этой пары с более высокой энергией удаляется из сгенерированного набора конформеров. Молекула, имеющая наименьшее число сгенерированных таким образом конформеров, считается молекулой сравнения, а ее конформеры также полагаются конформерами сравнения. Все конформеры оставшихся молекул набора сопоставляются с конформерами сравнения и осуществляется их фармакофорное отображение с обнаружением клики.

В рамках рассматриваемого алгоритма фармакофорные признаки для каждого конформера присваиваются автоматически. Отметим, что в этом случае расположение только гидрофобного признака (центра) совпадает с положением соответствующего гидрофобного атома, тогда как расположение остальных видов фармакофорных признаков берется иное. В частности, фармакофорный признак, соответствующий ароматическому кольцу, оказывается расположенным его центре, а фармакофорные центры, соответствующие Н-донорам и Н-акцепторам, расположены в точке, где должны находиться взаимодействующие с ними, соответственно, Н-акцепторы и Н-доноры сайта связывания белка. Такое расположение фармакофорных признаков позволяет при построении фармакофоров принимать во внимание ориентацию водородной связи и более корректно проводить сравнение фармакофоров.

Алгоритм фармакофорного отображения заключается в следующем. Для рассматриваемой пары конформеров определяются фармакофорные признаки и расстояния между ними. Для удобства понимания, обозначим фармакофорные признаки (например, H-доноры или H-акцепторы) одной молекулы как M_1, M_2,… , а для другой - m_1, m_2,… Далее рассматриваются все возможные сочетания однотипных фармакофорных признаков двух молекул, то есть H-доноров одной молекулы с H-донорами другой молекулы, H-акцепторов одной молекулы с H-акцепторами другой молекулы и так далее (M_1m_1, M_1m_2, M_2m_1, M_2m_2 …). На их основе создается *граф соответствия* (англ. correspondence graph), вершинами которого являются эти парные сочетания, причем вершины M_im_j и M_am_b связаны в этом графе ребром, если расстояние между фармакофорными признаками M_i и M_a в молекуле M отличается от расстояния между фармакофорными признаками m_j и m_b в молекуле m меньше, чем на заданную величину (Рис. 64). В общем случае граф соответствия может состоять из нескольких не связанных между собой компонент (Рис. 64).

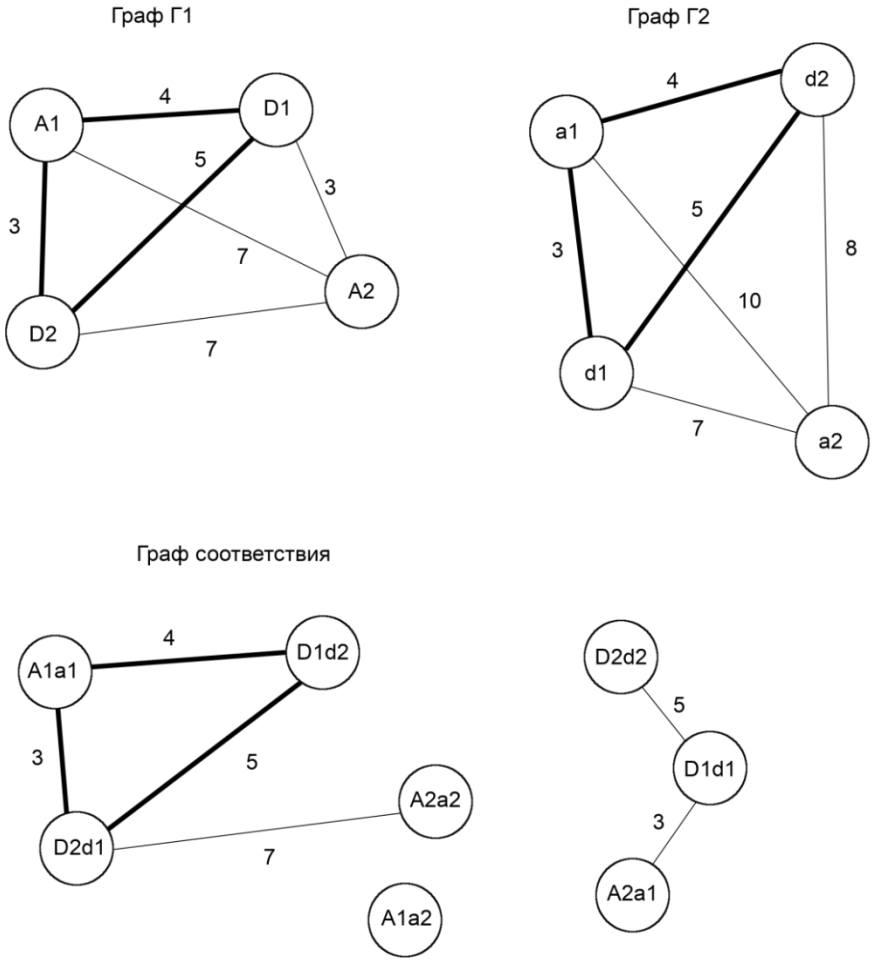


Рис. 64. Иллюстрация к формированию графа соответствия.

Наибольшие клики графа соответствия представляют собой наилучшие соотнесения между фармакофорными признаками двух молекул. Кроме того, и другие максимальные клики, которые, согласно приведенному выше определению, не являются частью друг друга, могут также быть использованы для построения фармакофорных моделей. Таким образом, после проведения фармакофорного отображения и построения максимальных клик путем сопоставления конформеров молекулы сравнения с конформерами всех остальных молекул формируется первоначальный набор фармакофоров. Программой проводится анализ найденных максимальных клик и те из них, которые встречаются хотя бы в одном конформере всех (или большинства) молекул, считаются наилучшими фармакофорными моделями. Последняя версия программы также проводит ранжирование различных фармакофорных моделей, основываясь на числе участвующих молекул, числе фармакофорных признаков в модели и расстояниях между ними. Кроме того, существует возможность автоматического перебора различных параметров метода, таких как допустимый интервал для расстояний, максимальное принимаемое во внимание расстояние между фармакофорными признаками, виды рассматриваемых фармакофорных признаков, с целью построения оптимальной фармакофорной модели.

Недостатком этого подхода является большая зависимость от вводимых конформаций, поскольку не существует способа их отбора внутри самого метода. Для успешного построения фармакофоров конформеры должны быть достаточно разнообразными, чтобы представлять все конформационное пространство, и низкоэнергетическими, поскольку чем выше энергия конформации, тем меньше вероятность ее реализации. Конформационно-жесткие молекулы являются наиболее удобными для данного способа фармакофорного отображения. В целом же качество фармакофорных моделей имеет тенденцию к снижению при использовании более конформационно-лабильных молекул.

4.2.4. Фармакофорное отображение с использованием метода максимального сходства

Алгоритм HipHop [181], реализованный в настоящее время в программе DS Catalyst Hypothesis, доступной как часть среды моделирования Discovery Studio [182] компании Accelrys, является наиболее известным алгоритмом фармакофорного отображения с использованием метода максимального сходства.

Так же, как и для DISCO (см. раздел 4.2.3), для начала работы алгоритма HipHop требуется осуществить предварительную генерацию наборов разнообразных конформеров, обладающих низкой энергией. Генерацию ведут с использованием алгоритма полинга [183, 184], сущность которого заключается в добавлении к функции энергии молекулы штрафа за сходство с уже сгенерированными конформерами.

Метод HipHop использует 5 типов фармакофорных признаков (H-акцептор, H-донор, центр положительного и отрицательного заряда, гидрофобный атом). Положение фармакофорных признаков определяется так же, как и в программе DISCO, то есть центры водородной связи полагаются размещенными в тех положениях, где должны находиться атомы белка, образующие водородную связь с соответствующими атомами лиганда.

На первом шаге программы определяются все общие фармакофоры молекул. Для этого генерируются все возможные наборы фармакофорных признаков, присутствующих в каждой молекуле, и для каждого набора рассчитываются расстояния между ними. Далее по всем молекулам производится поиск общих фармакофоров, которые реализуются хотя бы в одной конформации молекул. Считается, что фармакофор является для двух молекул общим, если выполняется 2 условия: (i) число и типы фармакофорных признаков в этих молекулах совпадают, (ii) расстояния между фармакофорными признаками в обоих фармакофорах совпадают в пределах заданной погрешности (см. ниже). В идеале хотелось бы найти фармакофор, который встречается во всех активных соединениях.

Следует отметить, что поскольку различные молекулы могут связываться с белком различным образом, одному сайту связывания белка могут соответствовать различные фармакофоры. Вместе с тем не все фармакофоры одинаково эффективны при скрининге. Например, фармакофоры, содержащие относительно редкие признаки, такие как положительный и отрицательный заряды, позволяют лучше разделять соединения с разной активностью при скрининге, чем фармакофоры, включающие часто встречающиеся и слабо взаимодействующие с белком гидрофобные центры. Чтобы принять эти обстоятельства во внимание, в алгоритме HipHop проводится ранжирование фармакофоров в соответствии со следующей скоринг-функцией:

$$score = M \sum_{x=0}^{K+1} q(x) \log_2 \left(\frac{q(x)}{p(x)} \right), \qquad (51)$$

где *M* – число активных соединений в выборке (поскольку выборка, как правило, состоит только из активных соединений, оно равно общему числу соединений в выборке), *K* – число фармакофорных признаков в данном фармакофоре (называемом в данном алгоритме гипотезой), *x* – класс гипотезы. В данном алгоритме выделяется *K+2* классов: если данная молекула полностью удовлетворяет фармакофору (гипотезе), то она относится к классу *K+1*. Если молекула не удовлетворяет данному фармакофору, но удовлетворяет такому фармакофору, в котором удалили один из признаков, то она относится к классу от 1 до *K* (поскольку в фармакофоре, состоящем из *K* признаков, есть *K* способов удалить вершину, то фармакофоры без одной вершины образуют *K* классов). Если молекула не удовлетворяет фармакофору, состоящему ни из *K*, ни из *K-1* центров, то молекула относится к классу 0. Таким образом, все молекулы из выборки получают отнесение к тому или иному классу. Долю молекул, относящихся к каждому классу *x*, от общего числа молекул (то есть отношение числа молекул, относящихся к данному классу, к общему числу классов) характеризует величина *q(x)*. Величина *p(x)* характеризует «редкость» данного фармакофора и равна частоте встречаемости данного фармакофора в большом наборе из «обычных» молекул. Эта величина определяется заранее на наборе соединений, обладающих самой различной биологической активностью, и «вшита» в программу. Таким образом, фармакофоры, которые часто встречаются среди активных соединений и редко в «обычных» молекулах, получают более высокий ранг.

Кроме того, особенностью метода максимального сходства служит то, что положение фармакофорных признаков задается центрированными в определенных точках пространства сферическими областями [185]. Если все фармакофорные признаки одной молекулы попадают внутрь всех таких областей гипотезы, то считается, что молекула удовлетворяет гипотезе (Рис. 65). Радиусы сфер могут отличаться в зависимости от типа фармакофорного признака или пожелания пользователя.

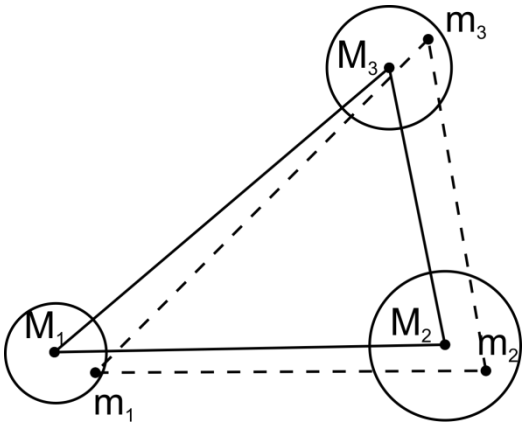


Рис. 65. Задаваемые сферами ограничения расположения фармакофорных признаков позволяют определить области пространства, в которой должны располагаться соответствующие фармакофорные признаки других молекул, чтобы фармакофор считался общим. Поскольку фармакофорные признаки m_1-m_2-m_3 находятся внутри ограничений, созданных вокруг фармакофорных признаков гипотезы M_1-M_2-M_3, фармакофор считается общим.

Похожий на HipHop алгоритм используется в программе Phase [186] от компании Schrödinger. На его первом шаге генерируются конформеры и с использованием правил, закодированных с помощью SMARTS (см. раздел 2.2.3.5 в пособии 1), определяются фармакофорные признаки. С использованием дерева поиска производится определение фармакофоров, которые являются общими либо для всей выборки активных соединений, либо для определенного их числа. Фармакофоры ранжируются на основании скоринг-функции, базирующейся на функции сходства фармакофоров. Последняя характеризует: (1) качество выравнивания молекул с данным фармакофором, (2) отклонение угла между направленными фармакофорными характеристиками (например, Н-связями), и (3) перекрывание объемов молекул.

Оригинальный метод фармакофорного отображения реализован в программе LigandScout [162, 187]. LigandScout, как и другие программы данного типа, работает с наборами предварительно сгенерированных конформеров, ранее это делалось с помощью программы OMEGA [188], в настоящее время имеется собственный генератор конформаций. Программа выделяет 6 основных типов фармакофорных признаков (включая липофильные и ароматические), причем в случае Н-акцепторов и Н-доноров рассматривается не только положение атома, но и направление образования водородной связи с ним. Имеется возможность использования фармакофорных признаков,

отвечающих за связывание с металлами в металлопротеинах, а также добавления иных центров с использованием SMARTS.

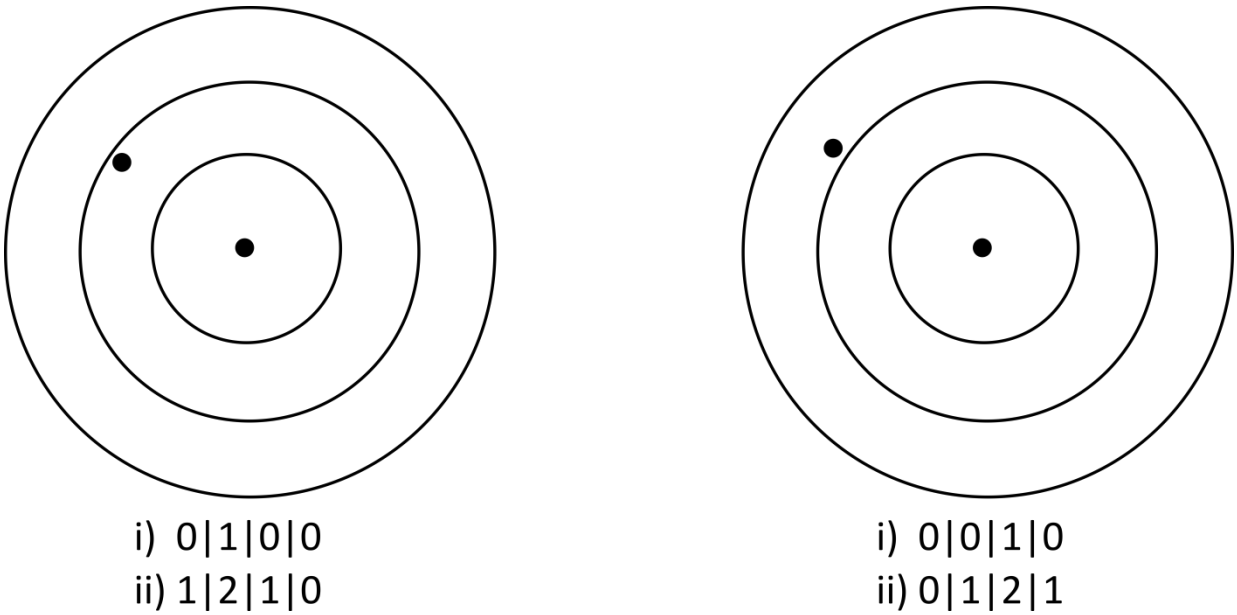


Рис. 66. Характеризация удаленности фармакофорных признаков в программе LigandScout с использованием корзин расстояний: (i) указание в какую корзину попадает фармакофорный признак, (ii) числа, которыми заполняется строка, после наложения биномиального распределения 1|2|1.

На первом этапе осуществляется выравнивание фармакофорных представлений двух молекул. Процесс выравнивания ведется в два этапа, на первом из которых осуществляется поиск оптимальных фармакофоров, а на втором – их выравнивание в пространстве. Поиск оптимальных фармакофоров осуществляется путем определения наиболее сходных групп фармакофорных признаков двух молекул. В отличие от других описанных выше подходов, расстояние между фармакофорными признаками в молекуле задается не с помощью геометрического расстояния, а с помощью так называемых *корзин расстояний* (англ. distance bin). В этом случае, если для заданной пары типов фармакофорных признаков (например, H-донор – липофильный атом) соответствующая пара фармакофорных признаков в молекуле расположена в интервале расстояний, характерном для данной корзины (например, 5.0 – 6.0 Å), то в данной корзине ставится 2, а в двух соседних ставится 1 (см. Рис. 66, где числа 1|2|1 соответствуют биномиальному распределению). Если в данную или соседние корзины расстояний попадает несколько пар однотипных признаков, то вклады

от них суммируются. Наложение «биномиального фильтра» сделано для того, чтобы не отклонять совпадения фармакофорных признаков двух молекул, если расстояние между фармакофорными признаками не попадает точно в данную корзину, а находится на почти границе с ней, но в другой корзине. Таким образом, для каждого фармакофорного признака молекулы создается T строк (T – общее число рассматриваемых видов фармакофорных признаков) длиной K (K – общее число корзин).

Для каждой пары молекул вычисляется матрица сходства, называемая также *матрицей затрат* (англ. cost matrix). Для двух молекул, содержащих N и M фармакофорных признаков соответственно, матрица сходства имеет размерность $N{\cdot}M$ и каждый ее элемент $c_{i,j}$ вычисляется следующим образом (для удобства полагаем, что в обеих молекулах и для всех фармакофорных признаков число корзин расстояний одинаково):

$$c_{i,j} = \sum_{k=1}^{K}\left(\omega(k)\cdot\sum_{t=1}^{T}\min\left(B_{i,t}(k), B_{j,t}(k)\right)\right), \qquad (52)$$

где первое суммирование идет по всем элементам строк, характеризующих расстояние между фармакофорными признаками типа t, второе суммирование идет по всем типам фармакофорных признаков; величина $B_{i,t}(k)$ - это значение, находящееся в k-м элементе строки, характеризующей расстояние от данного фармакофорного признака i до фармакофорных признаков типа t данной молекулы; $B_{j,t}(k)$ – то же, но для фармакофорного признака другой молекулы; min(A,B) – минимальное число из вариантов A и B. $\omega(k)$ – вектор весов, который компенсирует потерю информации при переходе от трехмерного представления к одномерному (корзинам расстояний). Кроме того, он дает определенное предпочтение более близко расположенным фармакофорным признакам по сравнению с удаленными.

С использованием *венгерского алгоритма*[1] (*Hungarian Matcher*) [189, 190] на основе модифицированной матрицы сходства

[1] Венгерский алгоритм (известный также как алгоритм Манкреса или Куна-Манкреса) – алгоритм оптимизации для решения задачи о назначениях. Суть задачи оптимального назначения – поиск такого назначения N видов работ N работникам, чтобы общая стоимость работ была минимальной (стоимость проведения каждой работы каждым работником записана в матрице). Используется в самых различных областях математики

осуществляется поиск оптимальных пар фармакофорных признаков и проводится их выравнивание с использованием алгоритма Кабша[1] [191, 192]. В некоторых случаях это невозможно сделать с заданной точностью, поскольку при переходе от трехмерного представления фармакофоров к их описанию с использованием корзин расстояний теряется информация. Тогда пара признаков, ответственная за ложное совпадение, исключается из рассмотрения, и поиск наилучших сочетаний фармакофоров продолжается. На завершающем этапе проверяется, удовлетворяет ли найденный фармакофор дополнительным ограничениям (таким как расположение векторов направлений водородных связей).

Объединение полученных фармакофоров с использованием выравнивания позволяет определить набор фармакофорных моделей и осуществить их ранжирование с помощью специальной скоринг-функции. В настоящее время в программе реализовано несколько скоринг-функций, зависящих от совпадения фармакофоров, качества наложения молекул, которым соответствуют данные фармакофоры (перекрывания объемов и т.п.). Базовой является функция фармакофорного совпадения, равная $score = (c \cdot n) + (9 - 3 \cdot \min(R, 3))$, в которой n – число совпавших в результате выравнивания фармакофорных признаков, R – среднеквадратичное отклонение фармакофорных признаков двух молекул, с – заданный пользователем коэффициент, характеризующей важность совпадения фармакофорных признаков (обычно равен 10). Эта функция стремится отбирать в первую очередь фармакофоры с максимальным числом совпавших фармакофорных признаков, и во вторую – те из них, для которых малы геометрические отклонения.

Программа LigandScout позволяет достаточно быстро проводить поиск оптимального фармакофора для нескольких десятков молекул, проводить с его помощью виртуальный скрининг контрольного набора молекул и уточнять параметры для создания более точной модели.

Основные недостатки методов фармакофорного отображения с использованием поиска по сходству являются следствиями

и программирования. См. подробнее https://ru.wikipedia.org/wiki/венгерский_алгоритм

[1] Алгоритм Кабша – алгоритм для выравнивания в 3D пространстве двух объектов, представленных набором вершин, так, чтобы среднеквадратичное расстояние между вершинами было наименьшим. Вершины двух объектов должны быть предварительно отнесены друг к другу.

необходимости генерации представительного набора конформеров с низкой энергией. Они почти такие же, как в методе DISCO: качество модели во многом зависит от того, насколько хорошо и репрезентативно отобраны конформеры для проведения скрининга. Чем более конформационно жесткие молекулы подвергаются анализу, тем более качественной будет фармакофорная модель.

4.3. ТОПОЛОГИЧЕСКИЕ ФАРМАКОФОРЫ

Топологические (двумерные) фармакофоры концептуально похожи на описанные выше трехмерные фармакофоры. Предполагается, что в соединениях, обладающих определенным видом биологической активности, топологические расстояния между выбранными фармакофорными признаками должны лежать в определенных интервалах. Набор таких фармакофорных признаков и топологических расстояний между ними формирует *фармакофор-запрос* (англ. query). В ходе двумерного фармакофорного поиска из базы данных отбираются соединения, удовлетворяющие этому запросу. Иначе говоря, задача двумерного фармакофорного поиска сводится к задаче установления изоморфного вложения графа фармакофора-запроса (называемого просто фармакофором) в молекулярный граф, в котором атомы заменены на соответствующие им фармакофорные признаки. В простейших случаях для спецификации фармакофора-запроса может быть использовано линейное представление SMARTS (см. раздел 2.2.3.5 в пособии 1).

Фармакофор запроса может быть выведен либо из экспериментальных данных о строении комплекса лиганда с сайтом связывания биологической мишени, либо путем сравнения структур активных соединений. В качестве примера можно привести двухточечный топологический фармакофор, состоящий центров положительного отрицательного заряда (Рис. 67), который был использован в работе [193] в качестве одного из быстрых фильтров для отсеивания неактивных соединений. Для ускорения поиска геометрические расстояния между атомами были заменены топологическими.

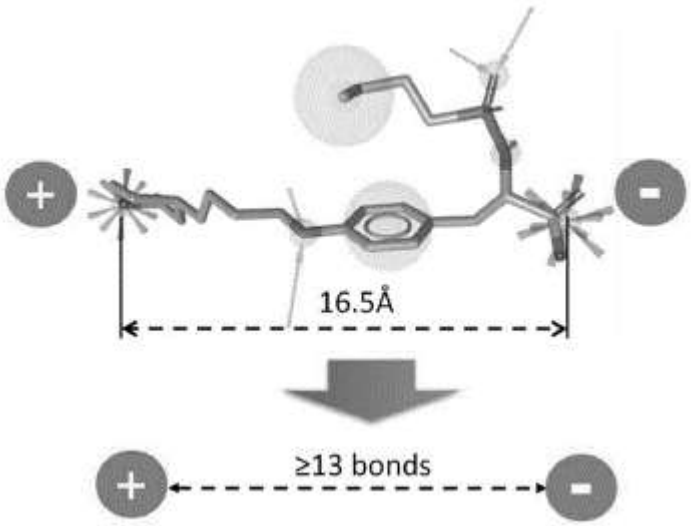


Рис. 67. Топологический и трехмерный фармакофоры, используемые при поиске антагонистов тромбоксановых рецепторов.

5. ВИРТУАЛЬНЫЙ СКРИНИНГ

5.1. КОНЦЕПЦИЯ ВИРТУАЛЬНОГО СКРИНИНГА

Под *виртуальным скринигом* понимается «просеивание» больших библиотек химических соединений (баз данных) через вычислительные «сита» (фильтры), приводящее к отбрасыванию соединений, предположительно «неактивных» либо обладающих неблагоприятными свойствами, и к формированию небольших наборов предположительно «активных» соединений, обогащенных желаемыми свойствами. Виртуальный скрининг, проводимый для поиска новых лекарственных средств, может быть основан либо на знании строения активных лигандов (англ. ligand-based), либо на знании строения макромолекул биологических мишеней (англ. structure-based). В настоящем пособии мы будем рассматривать лишь первый из этих двух видом виртуального скрининга. Хотя в настоящее время основной областью применения виртуального скрининга является разработка лекарственных средств, однако в последнее время он все чаще начинает использоваться при разработке новых материалов, катализаторов, комплексонов и других видов химических веществ.

5.1.1. Место виртуального скрининга в разработке лекарственных средств

Обычно считается, что основанный на виртуальном скрининге процесс разработки лекарственных препаратов включает 3 основных этап, Рис. 68:

1. *Первичный скрининг* (англ. *hit detection*) больших баз данных разнообразных органических соединений с целью нахождения нескольких перспективных молекул (хитов, англ. *hits*) для дальнейшей разработки.
2. *Оптимизация лидеров* (англ. *lead optimization*) с целью получения патентоспособных соединений, обладающих оптимальным набором фармакодинамических и фармакокинетических свойств, отсутствием токсичности и неблагоприятных побочных действий.
3. *Окончательная разработка* (англ. *development*), включающая доклинические и клинические испытания, формирование лекарственных форм и разработка технологии их производства.

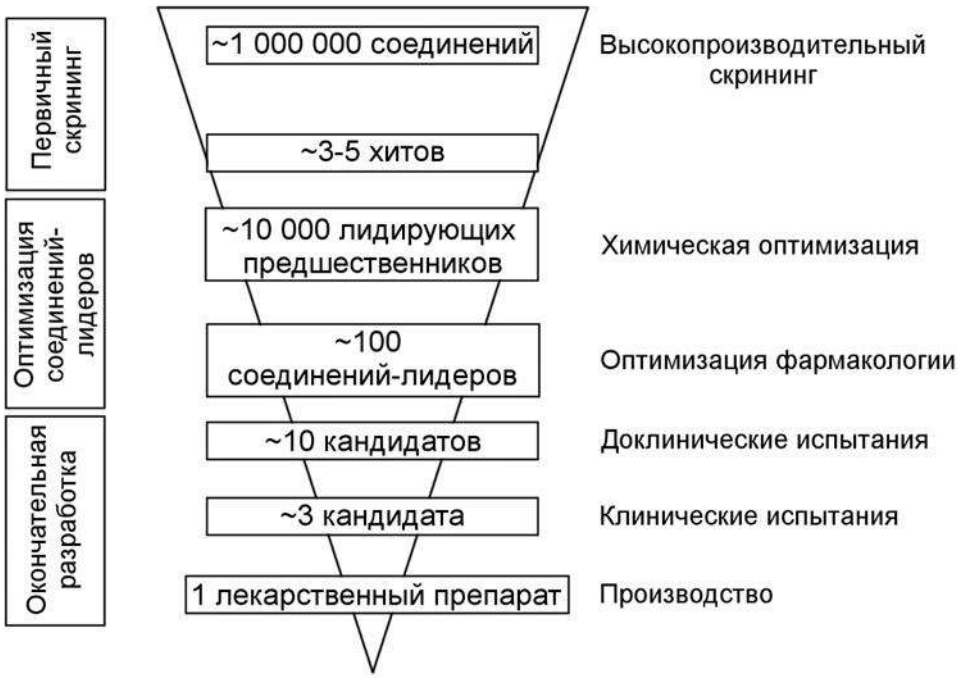


Рис. 68. Основные этапы разработки лекарственных препаратов.

Виртуальный скрининг может использоваться на первом и на втором этапах. Если на первом этапе скринингу обычно подвергаются большие библиотеки разнообразных химических соединений, то на втором этапе чаще используются сфокусированные комбинаторные библиотеки, созданные на основе соединений, найденных на первом этапе.

5.1.2. Воронка виртуального скрининга и ее компоненты

Виртуальный скрининг обычно представляют в виде *воронки* (англ. funnel), на вход которой поступает виртуальная библиотека, состоящая из большого числа молекул, а на выходе остается лишь небольшое число хитов. Такая воронка состоит из множества *фильтров*, задача каждого из которых заключается в отсеивании неперспективных молекул. Как правило, более производительные фильтры (позволяющие обрабатывать большое число соединений в единицу времени) предшествуют менее производительным (см. Рис. 69. Фильтры могут быть как *отсекающими* (принимающими решение, какое соединение отсечь), так и *ранжирующими* (оценивающими «качество» соединения при помощи непрерывной меры – оценки, англ. *score*). В последнем случае отсечение ведется путем ранжирования соединений по значению оценки и сравнением с выбранным пороговым значением. Примерами таких оценок являются:

- коэффициенты Танимото для фильтров, оценивающих сходство с активными соединениями;
- вероятность обладания определенным видом активности для фильтров, основанных на двухклассовых классификационных моделях «структура-свойство»;
- численное значение активности (pIC_{50}, pEC_{50} и др.) для фильтров, основанных на регрессионных моделях «структура-свойство» (QSAR);
- значение свободной энергии взаимодействия «лиганд-белок» либо ее аппроксимирующей скоринг-функции для фильтров, основанных на докинге.

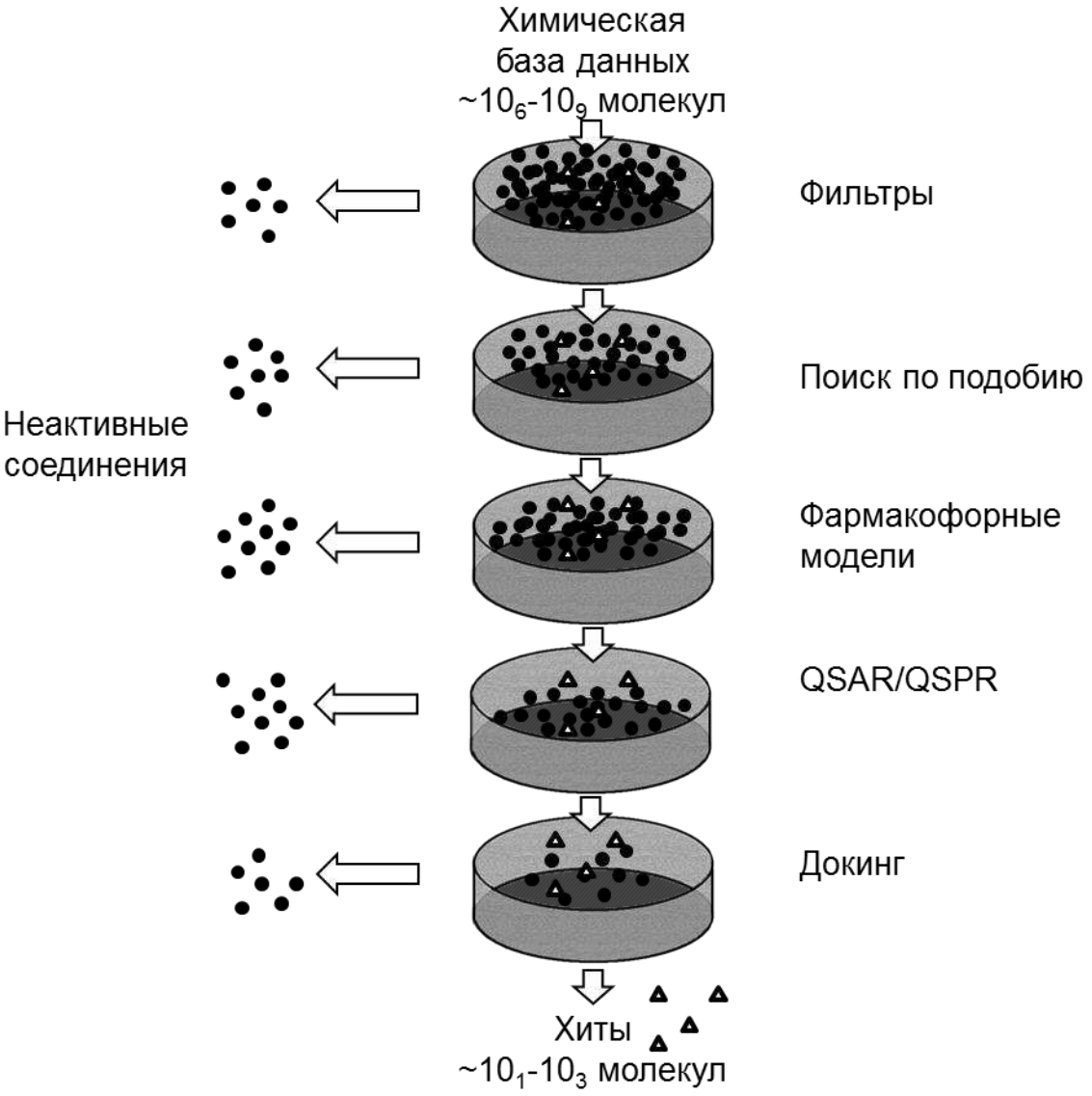


Рис. 69. Воронка виртуального скрининга.

Основные группы фильтров в совокупности с дополнительными вычислительными процедурами, необходимыми для их применения, составляют основные этапы виртуального скрининга:

1. Выбор или генерация библиотеки химических соединений для скрининга
2. Применение множества простых фильтров, ведущих к отсечению соединений по токсичности, биодоступности, стабильности молекул, «*лекарствоподобности*» (англ. *drug-likeness*), синтезируемости и ряду других критериев. Такие фильтры основаны либо на простых правилах, основанных на оценке физико-химических характеристик (например, правило Липинского), либо на поиске *алертов* (англ. *alerts*) – подструктур, наличие которых в молекуле может сигнализировать о ее «негодности» для использования в качестве лекарств;
3. Ранжирование с использованием 2D-структур (структурных формул):
    a. ранжирование по сходству с активными соединениями (например, при помощи индекса Танимото для сравнения «молекулярных отпечатков»). Этот компонент также называют *виртуальным скринингом на основе молекулярного подобия* (англ. similarity-based virtual screening, *SBVS*);
    b. ранжирование по склонности к обладанию нужным видом активности, оцениваемой при помощи одноклассовых моделей;
    c. ранжирование по вероятности обладания нужным видом активности, оцениваемой при помощи двухклассовых классификационных SAR моделей;
    d. ранжирование по численному значению активности, оцениваемой при помощи регрессионных моделей 2D QSAR.
4. Перевод структур из 2D в 3D, генерация представительного набора конформаций молекул.
5. Фильтрация с использованием 3D-структур:
    a. ранжирование по 3D-сходству с активными соединениями;
    b. фильтрация при помощи фармакофорных моделей;
    c. ранжирование по численному значению активности, оцениваемой при помощи регрессионных моделей 3D QSAR.
6. Докинг с использованием модели биологических мишеней

7. Молекулярная динамика и оценка с ее помощью свободной энергии связывания лигандов с биологическими мишенями (FEP, PB/SA, GB/SA)

Последние пункты 6 и 7 требуют знания структуры биологической мишени и в настоящем пособии не рассматриваются.

5.1.3. Числовые характеристики производительности компонент виртуального скрининга

При осуществлении виртуального скрининга с помощью ранжирующих фильтров решается задача *ранжирования* (англ. ranking) химических соединений в соответствии со значениями вычисленной для них оценки (скоринга). Хотя задача ранжирования может быть решена при помощи регрессионных либо классификационных моделей, принципиальным ее отличием от последних являются критерии оценки качества моделей. Для ранговых моделей важно не то, насколько точно они оценивают свойства химических соединений, а насколько правильно они позволяют их упорядочить по убыванию (либо возрастанию) активности. Хотя точный количественный прогноз величины биологической активности всегда ведет к правильному упорядочиванию по активности, однако оно может быть достигнуто и без точного количественного прогноза. На Рис. 70 приведен пример идеального ранжирования для набора из 11 соединений. Наверху приведены значения вычисленной для них оценки (скоринга), приписывающей более высокие значения более активным соединениям. Значения для соединений, для которых известно, что они активные, отмечены подчеркиванием. Снизу приведены значения этих же оценок, отсортированные по уменьшению значения. Видно, что активные соединения расположены в нем левее (до) неактивных, что и является целью идеального ранжирования. Еще ниже приведены значения вычисленных для них *рангов* – натуральных чисел от 1 до 11. При идеальном ранжировании ранги активных соединений должны быть меньше рангов неактивных. Наконец, в нижней строке приведены нормализованные ранги, приведенные к интервалу от 0 до 1.

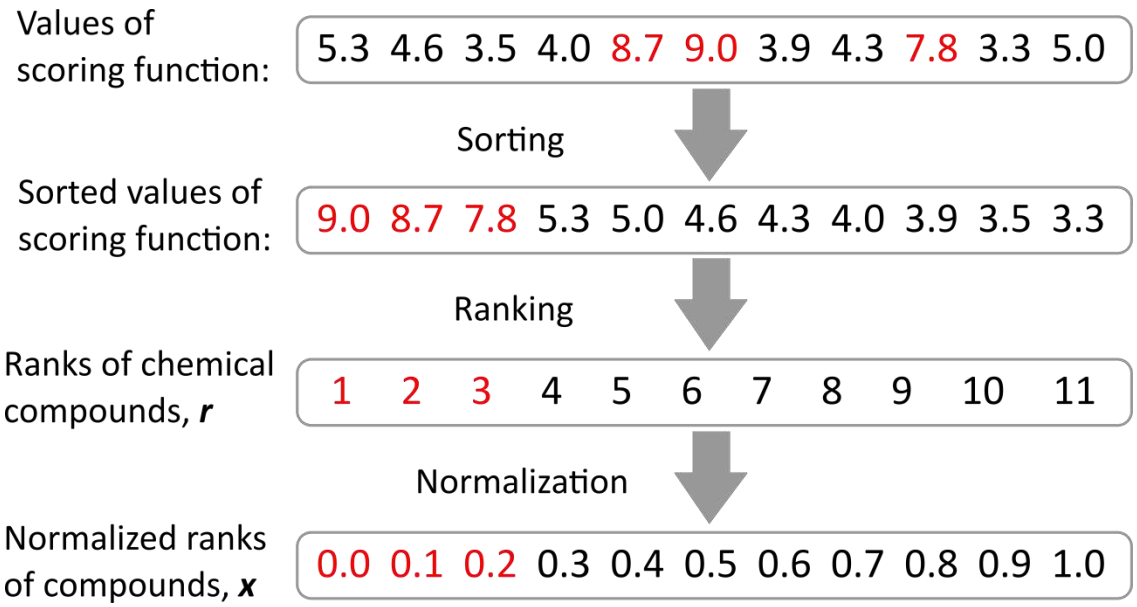


Рис. 70. Пример вычисления обычных и нормализованных рангов для выборки из 11 соединений, исходя из вычисленных для них значений оценочной функции.

Количественный показатель качества ранжирования с помощью ранговой модели называется ее *метрикой*. Существует два основных типа метрик для используемых в виртуальном скрининге ранговых моделей: метрики «*среднего распознавания*» (англ. average recognition) и метрики «*раннего распознавания*» (англ. early recognition) [194].

5.1.3.1. Метрики «среднего распознавания»

Метрики «среднего распознавания» вычисляют усредненный показатель с равным вкладом от всех активных соединений. Простейшей из таких метрик является *средний относительный ранг активных* <*x*>, равный среднему значению нормализованного ранга для активных соединений:

$$< x >= \frac{1}{N_a} \sum_{i=1}^{N_a} x_i \qquad (53)$$

где x_i – нормализованный ранг i-ого активного соединения, N_a – число активных соединений, а суммирование ведется только по активным соединениям. Значения <*x*> всегда больше нуля, но меньше единицы. Чем ближе к нулю значение <*x*>, тем лучше соответствующий ранжирующий фильтр. Для очень плохого фильтра, для которого положение активных соединений никак не связано с их рангами, значение <x> близко к 0.5. Значения в интервале от 0.5 до 1 соответствуют ранжирующим фильтрам, проводящим упорядочение в

обратном направлении, когда неактивные соединения предшествуют активным. На Рис. 71 приведены примеры хорошего фильтра (сверху) с <*x*> = 0.1 и плохого фильтра (внизу) с <*x*> = 0.5 для приведенного на Рис. 70 набора из 11 соединений. Для плохого фильтра активные соединения, отмеченные подчеркиванием и курсивом, равномерно распределены в списке нормализованных рангов.

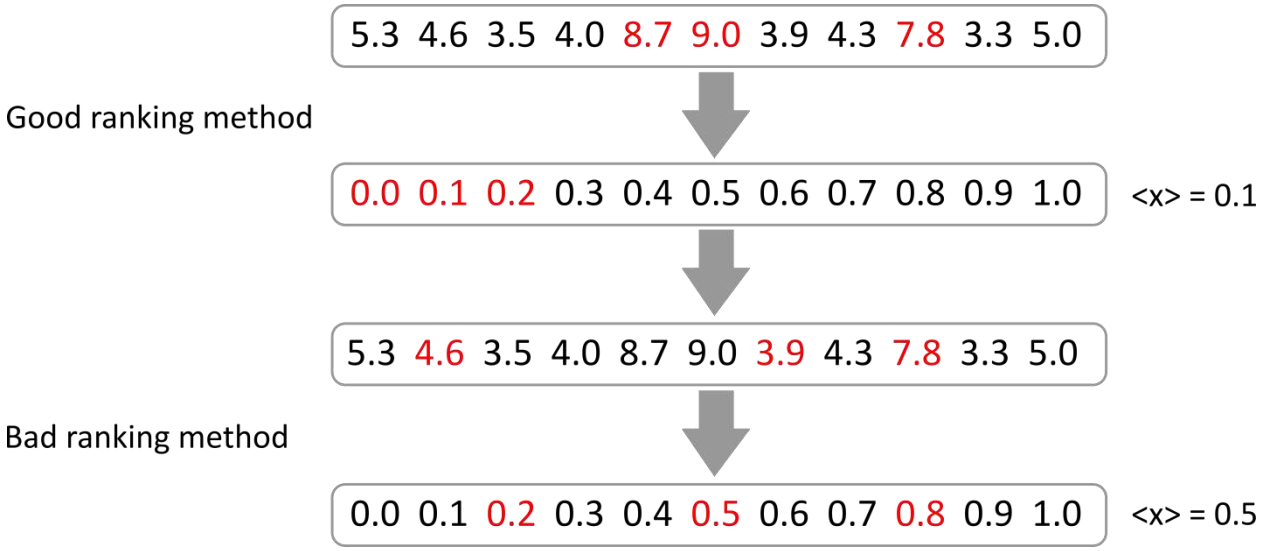


Рис. 71. Примеры вычисления <*x*> для хорошего (сверху) и плохого (снизу) ранжирования при помощи фильтра

Еще одной метрикой «среднего распознавания» для ранжирующих фильтров является *площадь под кривой накопления* (англ. Area Under Accumulation Curve, *AUAC*), которая может быть найдена путем интегрирования *кривой накопления* $F_a(x)$:

$$AUAC = \int_0^1 F_a(x)dx \qquad (54)$$

где x – нормализованный ранг химического соединения, когда набор соединений отсортирован от наилучшего к наихудшему, $F_a(x)$ – кумулятивная доля активных соединений с нормализованным рангом, не превышающим x. Статистический смысл $F_a(x)$ – это вероятность того, что у активного соединения нормализованный ранг не превышает x. Статистический смысл *AUAC* – это вероятность того, что у активного соединения, извлеченного из распределения с кумулятивной функцией $F_a(x)$, ранг окажется меньше ранга случайно выбранного соединения.

На Рис. 72 приведены примеры кривых накопления для ранжирующих фильтров хорошего и плохого качества. Для хороших фильтров AUAC приближается к единице, а для плохих – близко к 0.5. При вычислении значений определенных интегралов по методу

трапеций минимальное и максимальное возможные значения AUAC зависят от доли активных соединений в выборке:

$$AUAC_{min} = N_a/2N_t \quad (55)$$

$$AUAC_{max} = 1 - N_a/2N_t \quad (56)$$

где N_a – число активных соединений, N_t – общее число соединений в тестируемой выборке.

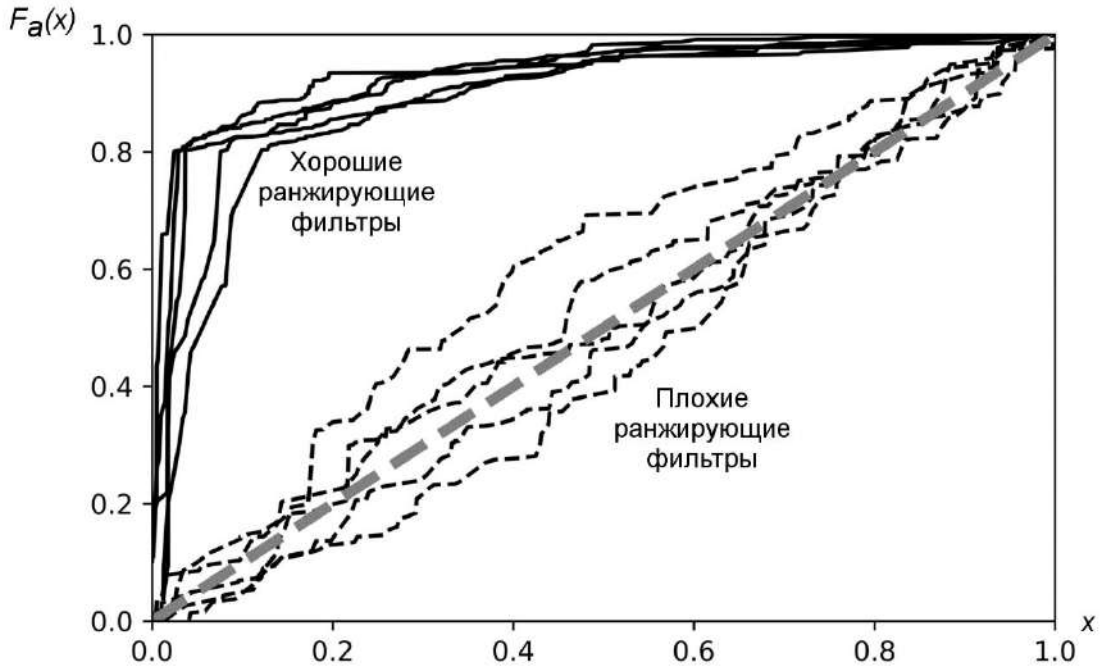


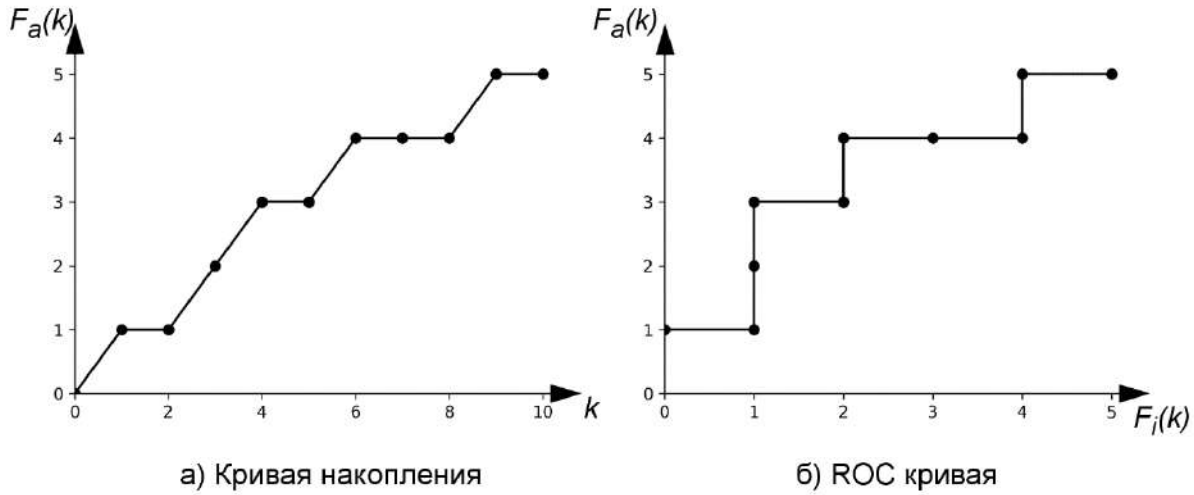
Рис. 72. Кривые накопления

Величины AUAC и <x> взаимосвязаны:

$$AUAC = 1 - \langle x \rangle \quad (57)$$

Наконец, третья и самая популярная метрика для ранжирующих фильтров – это AUC, площадь под ROC кривой, подробно рассмотренная в разделе 2.3.3 Пособия 3. Легко можно заметить сходство между *кривыми накопления* (англ. Accumulation Curve, *AC*) и кривыми ROC: все они при движении слева направо могут либо возрастать, либо идти горизонтально, но не снижаться, и все они начинаются в левом нижнем углу и заканчиваются в правом верхнем. Также во всех случаях площади под ними характеризуют качество работы фильтра, причем очень хорошие фильтры имеют значения AUAC и AUC, близкие к единице, а плохие – близкие к 0.5. Тем не менее, между кривыми AC и ROC и, соответственно, между AUAC и AUC имеются существенные различия. Для пояснения рассмотрим

Рис. 73, на котором представлены кривые AC и ROC для набора из 10 соединений, пять из которых (с рангами 1, 3, 4, 6, 9) являются активными.



а) Кривая накопления б) ROC кривая

Рис. 73. Сравнение кривой накопления (слева) с кривой ROC (справа).

На Рис. 73 k – ранг соединения в соответствии с осуществляемым фильтром ранжированием, $F_a(k)$ – это кумулятивная доля активных соединений с нормализованным рангом, не превышающим x, $F_i(k)$ –это аналогично построенная кумулятивная доля неактивных соединений. Видно, что в кривых ROC могут быть вертикальные отрезки. Основное же преимущество кривых ROC над кривыми накопления заключается в том, что AUC для идеальной модели в точности равен 1, тогда как значение AUAC для идеальной модели меньше 1 на величину, зависящую от доли активных соединений в выборке. Благодаря этому, кривыми ROC удобнее пользоваться, чем кривыми накопления, поскольку, исходя из значений AUC, всегда можно сказать, является ли ранжирующий фильтр идеальным. Кроме того, величина AUC обладает очень четким статистическим смыслом – это вероятность того, что ранг произвольного активного соединения меньше ранга произвольного неактивного.

Можно показать [194], что существует взаимосвязь между значениями тех рассмотренных выше метрик «среднего распознавания»:

$$AUC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i} = \frac{1 - < x >}{R_i} - \frac{R_a}{2R_i} \qquad (58)$$

где R_a – доля активных соединений, а R_i – доля неактивных соединений.

5.1.3.2. Метрики «раннего распознавания»

Хотя метрики «среднего распознавания», такие как площадь под ROC-кривой (AUC), хорошо характеризуют способность ранжирующих фильтров отделять активные от неактивных соединений, тем не менее, это не совсем то, что должно характеризовать качество процедуры виртуального скрининга. Ключевым критерием успеха ранжирующего фильтра в виртуальном скрининге является способность расположить активные соединения (пусть даже и не все) в самом начале упорядоченного списка соединений, поскольку лишь небольшая часть соединений будет отобрана. В то же время способность фильтра правильно ранжировать по активности отбрасываемые соединения совершенно не влияет на его практическую ценность. Вследствие этого метрики «среднего распознавания», не делающие различие между способностью корректно ранжировать отбираемые и отбрасываемые в процессе виртуального скрининга соединений, недостаточно хорошо характеризуют практическую ценность фильтров.

Рис. 74. Проблема «раннего распознавания»: три ранжирующих фильтра A, B, C, которые характеризуются одинаковыми значениями метрик «среднего распознавания», приводят к разному числу отбираемых активных соединений.

Для пояснения вышеизложенной проблемы, на Рис. 74 представлены 3 варианта ранжирования выборки из 30 соединений, 6 из которых активны (обозначены единицами), а 24 – неактивны (обозначены нулями). Предполагается, что 3 наилучших соединения будут отобраны, а остальные 27 – отброшены. Все три варианта характеризуются одинаковыми значениями метрик «среднего распознавания». Тем не менее, в варианте A будут отобраны 3 активных соединения, в варианте B – одно, а в варианте C вообще не будут отобраны активные соединения. Следовательно, одних метрик

«среднего распознавания» недостаточно для определения практической ценности ранжирующего фильтра для виртуального скрининга. В этом состоит суть проблемы «раннего распознавания».

Для решения проблемы «раннего распознавания» предложен целый ряд подходов. Самой простой метрикой «раннего распознавания» является *фактор обогащения* (англ. *Enrichment Factor*, $EF_{x\%}$), который показывает, во сколько раз чаще можно найти активные соединения в верхних $x\%$ упорядоченного по значению оценки (скоринга) списка по сравнению с долей активных соединений в тестируемом наборе данных (равной вероятности «вытянуть» активное соединения при случайном отборе из списка):

$$EF_{x\%} = \frac{Hits_{x\%}/N_{x\%}}{N_a/N_t} \qquad (59)$$

где $Hits_{x\%}$ – число активных соединений в верхних $x\%$ упорядоченного списка; $N_{x\%}$ – общее число соединений в верхних $x\%$ упорядоченного списка; N_a – число активных соединений в списке; N_t – общее число соединений в списке. Следует подчеркнуть, что фактор обогащения является параметрической метрикой, поскольку его значение зависит от параметра $x\%$. В качестве $x\%$ часто берут 1%, 5%, 20%.

Хотя фактор обогащения является популярной метрикой «раннего распознавания» с интуитивно понятным смыслом, у этой характеристики, тем не менее, есть ряд недостатков. Во-первых, значение этой метрики не зависит от того, насколько хорошо в отобранной доле соединений активные соединения отделены от неактивных. Во-вторых, то же самое относится к отбрасываемой доле (100-x)% соединений. В-третьих, при небольшом числе отбираемых соединений значение фактора обогащения очень сильно зависит от небольших вариаций в выборке. В-четвертых, минимальное и максимальное возможные значения фактора обогащения не являются константой и зависят от доли активных соединений в выборке.

Для преодоления первых трех недостатков было предложено заменить строгий отбор первых $x\%$ соединений (когда всем отбираемым соединениям придается одинаковый вес 1, а всем отбрасываемым – одинаковый вес 0) на взвешивание с помощью экспоненциальной функции (фактора Больцмана), которая придает первым соединениям в упорядоченном списке больший вес. На Рис. 75 проиллюстрирована идея использования экспоненциальной функции в качестве «нечеткой» альтернативы строгому отбору соединений. Внизу приведена строка, в которой активные соединения в упорядоченном списке обозначены символом «1», а неактивные – «0». Сверху приведен набор экспоненциальных кривых для разных

значений параметра α. На рисунке видно, что активным соединениям в начале списка придается намного больший вес, чем соединениям в конце списка. Кроме того, чем больше значение параметра α, тем меньшему числу соединений придается значительный вес. Поэтому считается, что величина 1/α имеет тот же смысл, что и процент отбираемых соединений *x*%, т.е. чем меньше значение 1/α, тем отбор меньшего числа соединений описывает «нечетким образом» экспоненциальная функция.

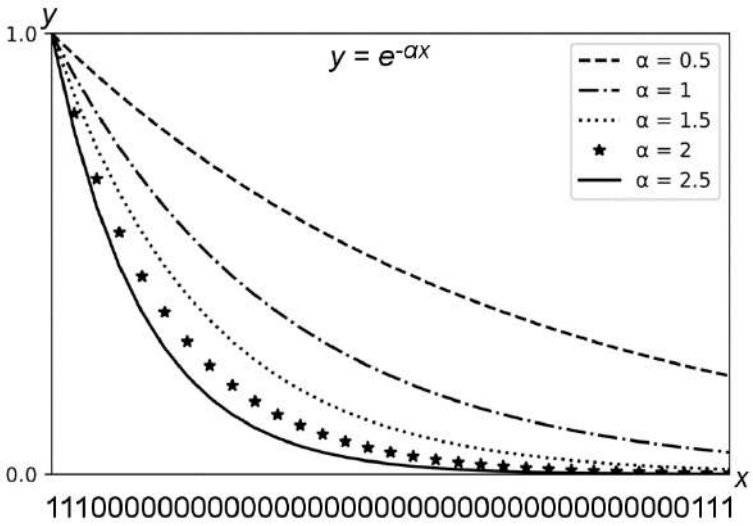


Рис. 75. Использование экспоненциальной функции как «нечеткой» альтернативы строгому отбору соединений

Выраженный при помощи экспоненциальной функции «нечеткий» аналог фактора обогащения получил название *робастного начального улучшения* (англ. Robust Initial Enhancement, RIE_α):

$$RIE_\alpha = \frac{\sum_{i=1}^{N_a} e^{-\alpha x_i}}{\langle \sum_{i=1}^{N_a} e^{-\alpha x_i} \rangle_{x \sim U(0,1)}} = \frac{\frac{1}{N_a}\sum_{i=1}^{N_a} e^{-\alpha x_i}}{\frac{1}{N_t}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N_t}-1}\right)} \qquad (60)$$

где $U(0,1)$ – равномерное распределение в интервале от 0 до 1, x_i – нормализованный ранг, а смысл N_a и N_t тот же, что и для фактора обогащения. Так же, как и фактор обогащения $EF_{x\%}$, RIE_α является параметрической величиной, но в качестве параметра используется α.

Хотя применение RIE_α в качестве метрики «раннего распознавания» позволяет преодолеть первые три из четырех указанных выше недостатка фактора обогащения, но четвертый недостаток остается: минимальное и максимальное значения зависят от доли активных соединений в выборке и значения параметра α:

$$RIE_{min} = \frac{1 - e^{\alpha N_a/N_t}}{(1 - e^{\alpha})\, N_a/N_t} \qquad (61)$$

$$RIE_{max} = \frac{1 - e^{-\alpha N_a/N_t}}{(1 - e^{-\alpha})\, N_a/N_t} \quad (62)$$

Для преодоления этого последнего недостатка была введена нормированная версия этой характеристики, получившая название *улучшенная по Больцману дискриминация ROC* (англ. *Boltzmann-Enhanced Discrimination of ROC, BEDROC$_\alpha$*), минимальное и максимальное значения которой равны, соответственно, 0 и 1:

$$BEDROC_\alpha = \frac{RIE_\alpha - RIE_{min}}{RIE_{max} - RIE_{min}} \quad (63)$$

Эта метрика «раннего распознавания» также зависит от параметра α. Статистический смысл *BEDROC$_\alpha$* состоит в том, что ее значение равно вероятности того, что ранг произвольного активного соединения окажется меньше ранга произвольного соединения, случайно выбранного в соответствии с экспоненциальным распределением с параметром α. Таким образом, *BEDROCα* является аналогом AUC ROC, приспособленным для решения проблемы «раннего распознавания»

5.2. БАЗЫ ДАННЫХ ДЛЯ ВИРТУАЛЬНОГО СКРИНИНГА

Ниже приведены базы данных большого размера, которые могут быть использованы при проведении виртуального скрининга.

- ZINC15 (http://zinc15.docking.org/). Свободная для скачивания база данных, содержащая структуры коммерчески доступных соединений (>130 миллионов соединений в 3D-формате, подготовленному к докингу).
- GDB-11, GDB-13 (~1 миллиард соединений), GDB-17 (~166 миллиардов соединений) (http://gdb.unibe.ch/downloads/). Все малые органические молекулы с числом неводородных атомов 11, 13 и 17, содержащие C, N, O, F, (S, Cl), удовлетворяющие правилам химической устойчивости и синтезируемости,
- SAVI (https://cactus.nci.nih.gov/download/savi_download/). ~300 миллионов соединений, синтезируемых в одну стадию из каталожных реактивов при помощи набора синтетических правил, описанных в экспертной системе по планированию органического синтеза LHASA.
- PGVL (*Pfizer Global Virtual Library*) – виртуальная база данных, построенная из возможных продуктов 1200 комбинаторных реакций, насчитывающая 10^{14}-10^{18} виртуальных соединений [195].

5.3. ПРОСТЕЙШИЕ ФИЛЬТРЫ ДЛЯ ВИРТУАЛЬНОГО СКРИНИНГА

Когда проводится виртуальный скрининг, для отбрасывания химических соединений, которые могут обладать неблагоприятными свойствами (такими как, например, плохая биодоступность, нестабильность, токсичность и др.) используют простейшие фильтры двух основных категорий: (1) правила на основе физико-химических характеристик и состава (числа тех или иных атомов и связей) молекул и (2) структурные алерты, показывающие присутствие в молекулах тех или иных сложных структурных фрагментов.

5.3.1. Правила биодоступности на основе физико-химических характеристик и состава молекул

Основанные на физико-химических характеристиках и составе молекул простейшие фильтры используются для отбрасывания соединений с плохой биодоступностью вследствие неблагоприятных для лекарственного действия процессов абсорбции в тканях человеческого организма и транспорта между ними (т.н. свойства ADME – *Absorption, Distribution, Metabolism, Excretion*). Кроме того, основанные на оценках состава молекул (числа тех или иных атомов и связей) простейшие фильтры также используются для отбрасывания химических соединений, которые сильно отличаются от большинства молекул лекарств.

5.3.1.1. Правила биодоступности

Наиболее известным правилом биодоступности является *правило Липинского* (т.н. *«правило пяти»*) для оценки возможности перорального (прием таблеток) введения химического соединения в организм [196, 197]. Согласно этому правилу:

- число доноров водородных связей (обычно вычисляется как суммарное число групп OH и NH) не должно превышать 5;
- молекулярная масса должна быть меньше 500;
- липофильность *log P*, вычисленная при помощи программы CLOGP, не должна превышать 5;
- число акцепторов водородных связей (обычно вычисляется как суммарное число атомов N и O) не должно превышать 10.

Если входящее в состав лекарственного препарата биологически-активное химическое соединение не удовлетворяет перечисленным

выше правилам, то, скорее всего, при пероральном применении такого препарата оно в организме у человека не дойдет до соответствующей биологической мишени и потому не окажет необходимого воздействия на организм. Всем четырем перечисленным выше правилам удовлетворяет 87% биологически-активных соединений, составляющих основу лекарственных препаратов. В то же время одному из этих правил не удовлетворяет 7% соединений, двум правилам – 4%, а трем правилам – только 1% соединений.

Кроме правила Липинского, существует ряд других правил биодоступности. В частности правила Гозе с соавт. [198] были созданы путем анализа 7183 соединений, входящих в состав базы Comprehensive Medicinal Chemistry (CMC). Согласно этим правилам чтобы молекула была биодоступна:

- молекулярная масса должна быть в интервале от 160 до 480;
- липофильность (оцениваемая с помощью метода ALOGP) должна быть в интервале от -0.4 до 5.6;
- молярная рефракция должна быть в интервале от 40 до 130;
- общее число атомов в молекуле должно быть в интервале от 20 до 70.

Всем перечисленным выше правил удовлетворяют более 80% молекул лекарств.

Правила Вебера с соавт. [199] устанавливают следующие критерии биологической доступности при пероральном введении лекарств:

- в молекуле должно быть меньше 10 свободно вращающихся связей;
- площадь полярной части поверхности молекулы (площади полярной поверхности, PSA) не должна превышать 140 $Å^2$.

Поскольку при разработке лекарственных препаратов на этапе оптимизации лидера для увеличения аффинности и селективности увеличивают число атомов (а, значит, и молекулярную массу) и липофильность молекул, то, чтобы не выйти за границы, определяемые «правилами пяти» Липинского, на первоначальном этапе скринирования для выявления потенциальных лидеров было предложено использовать «правила трех» [200] с заниженными верхними границами:

- липофильность молекул *log P* не должна превышать 3;
- молекулярная масса должна быть меньше 300;
- число доноров водородных связей не должно превышать 3;
- число акцепторов водородных связей не должно превышать 3;

- не должно быть больше 3 свободно вращающихся связей в молекуле.

5.3.1.2. Правила «лекарствоподобия»

При виртуальном скрининге обычно отбрасывают соединения, состав которых сильно отличается от состава типичных лекарственных препаратов. Часто отбрасывают соединения, которые:

- включают неподходящие катионы (например, тяжелые металлы) и анионы (например, цианид-ионы);
- с молекулярной массой меньшей 150 и большей 750 дальтон;
- содержащие химические элементы, не входящий в список (C, H, O, N, P, S, F, Cl, Br, I);
- с общим числом атомов меньше 10;
- в которых нет ни одной из связей из списка (C-C, C-N, C-O, C-S);
- можно вводить и иные правила.

5.3.1.3. Разработка правил биодоступности и «лекарствоподобия»

Разработка правил биодоступности и «лекарствоподобия» включает обычно три этапа. Прежде всего, выбирают ключевые физико-химические характеристики соединений для построения правил. Далее по каждой из таких характеристик строят гистограмму, показывающую частоту встречаемости в базе данных соединений, для которых значения выбранной характеристики находятся внутри заданных для столбцов гистограммы интервалов. После этого для каждой гистограммы определяют с помощью статистических расчетов квантили Q_α, для которых, согласно определению, вероятность того, что значение характеристики A не превышает Q_α, не меньше параметра α:

$$P(A \leq Q_\alpha) \geq \alpha \qquad (64)$$

В этом случае найденные значения квантилей определяют интервал значений характеристики A, в который попадают $100\times(1-2\alpha)$ процентов соединений.

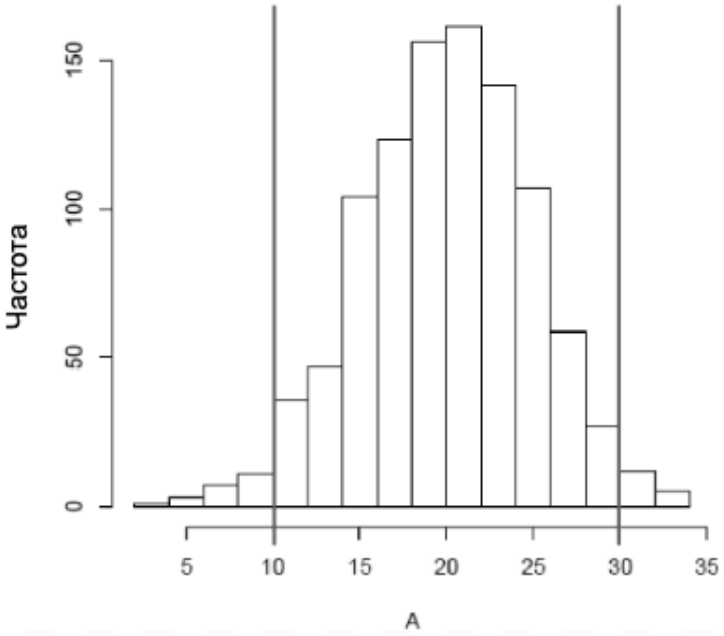


Рис. 76. Гистограмма для формирования эмпирического правила для распределения характеристики *A*.

В качестве примера на Рис. 76 приведена гистограмма, построенная для условной характеристики *A* на наборе из 1000 соединений. Пусть мы хотим построить правило, которому подчиняются 95% соединений. В этом случае значение параметра α будет (100-95)/2 = 0.025. Для этого значения параметра значения нижней и верхней квантилей для гистограммы будут:

$$Q_{\alpha}(A) = Q_{0.025}(A) = 10.3 \qquad (65а)$$

$$Q_{1-\alpha}(A) = Q_{0.975}(A) = 29.3 \qquad (65б)$$

Это приводит к «правилу» $10 < A < 30$, которому удовлетворяет около 95% соединений.

5.3.2. Структурные алерты

Структурными алертами называются подструктуры (фрагменты химических структур), наличие которых у химического соединения может свидетельствовать о его неблагоприятных свойствах, например, токсичности, высокой реакционной способности, нестабильности и др. Вследствие этого наличия структурных алертов в химических структурах лекарственных препаратов и химикатов, производимых химической промышленностью, следует избегать.

К основным категориям структурных алертов относятся алерты токсичности, «*лекарствоподобия*» (англ. drug likeness), высокой реакционной способности и нестабильности, PAINS, биоразлагаемости

и др. Часто один и тот же структурный алерт может быть отнесен сразу к нескольким категориям. Всего идентифицировано больше 2000 различных структурных алертов, наличие которых в химических структурах следует избегать, и поэтому их следует отбрасывать при виртуальном скрининге.

Структурные алерты удобно представлять с помощью строк SMARTS (см. раздел 2.2.3.5 в пособии 1), что дает возможность автоматизировать их хранение и применение в процессе виртуального скрининга для выявления их присутствия в молекулах химических соединений.

5.3.2.1. Структурные алерты токсичности

Структурными алертами токсичности (англ. *toxicity structural alerts*) являются подструктуры (фрагменты), наличие которых в молекулах химического соединения свидетельствует о возможности его неблагоприятного воздействия на органы человека. В некоторых публикациях их также называют *токсикофорами* (англ. *toxicophores*). С каждым структурным алертом токсичности, как правило, связан тот или иной механизм действия.

Ключевым процессом, определяющим действие большинства структурных алертов токсичности, является ковалентное связывание электрофильных *ксенобиотиков* (чужеродных для живых организмов химических веществ) с нуклеофильными группами (такими как $–NH_2$, -SH), содержащимися в образующих живые организмы макромолекулах белков либо нуклеиновых кислот. На Рис. 77 в качестве примера показана реакция Михаэля между электрофильным ксенобиотиком (акролеином) и тиоловой группой цистеинового аминокислотного остатка в макромолекуле белка.

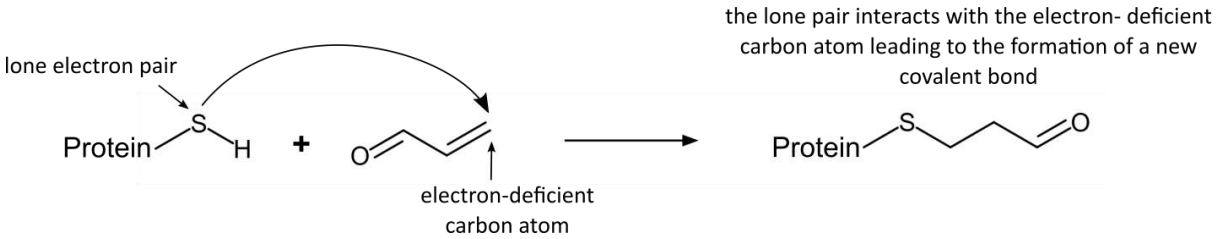


Рис. 77. Реакция Михаэля между электрофилом (акролеином) и тиоловой группой в макромолекуле белка

Участвующие в ковалентной модификации эндогенных биомолекул электрофильные химические соединения могут попадать в

живой организм в неизменном виде извне либо могут быть образованы в самом организме в результате метаболических реакций окисления либо фотоактивации. В первом случае структурные алерты идентифицируют функциональные группы с присущей им химической реакционной способностью (англ. *inherent chemical reactivity*), которые выступают в качестве электрофилов при ковалентном связывании с нуклеофильными группами биомолекул. Примеры таких алертов приведены на Рис. 78.

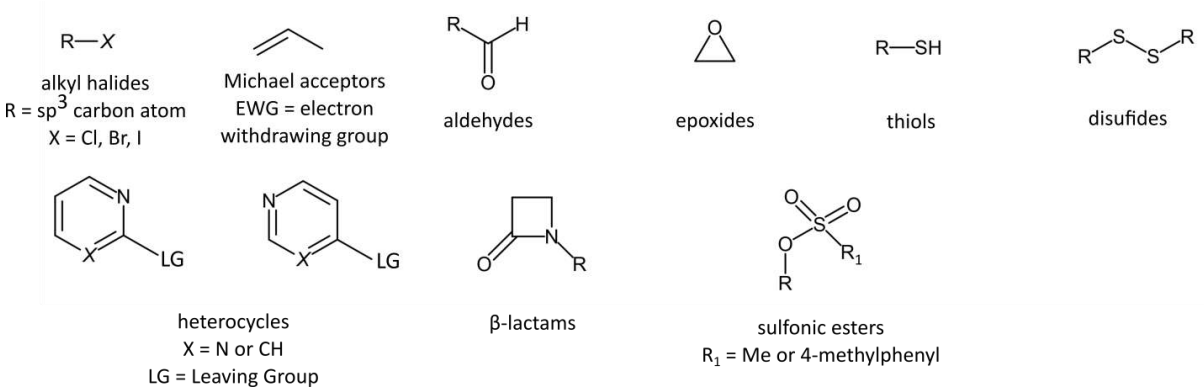


Рис. 78. Примеры структурных алертов, обозначающие функциональные группы с присущей им способностью выступать в качестве электрофилов при реакции с нуклеофильными группами биомолекул.

Во втором случае сами по себе молекулы ксенобиотиков не могут вступать в реакцию с нуклеофильными группами биомолекул, но приобретают такую способность в результате метаболической активации. Часто такая активация происходит в клетках печени путем окисления под воздействием ферментов группы цитохрома-450. В этом случае становится задействованным природный механизм выведения из организма гидрофобных соединений путем их гидрофилизации при помощи окисления. В качестве примера на Рис. 79 приведен механизм метаболической активации галогенобензолов путем окисления в клетках печени кислородом воздуха с помощью фермента цитохром-450 (CYP450). Хотя сами галогенобензолы не могут вступать в реакцию с биомолекулами, но в результате метаболической активации образуются электрофильные молекулы с очень высокой реакционной способностью по отношению к нуклеофильным группам биомолекул.

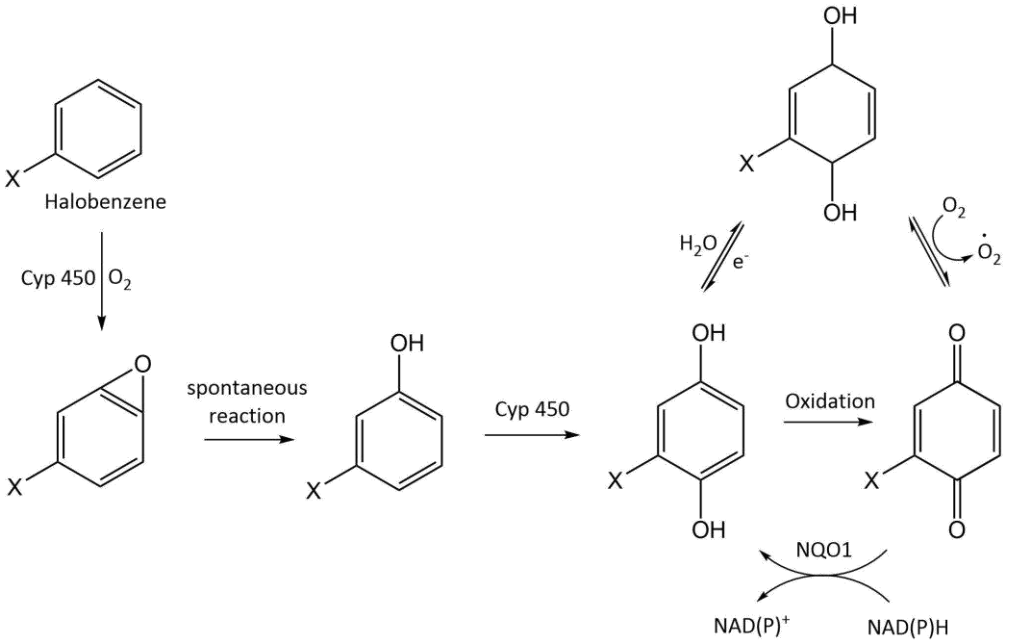


Рис. 79. Механизм метаболической активации галобензолов при окислении в клетках печени с помощью фермента CYP450

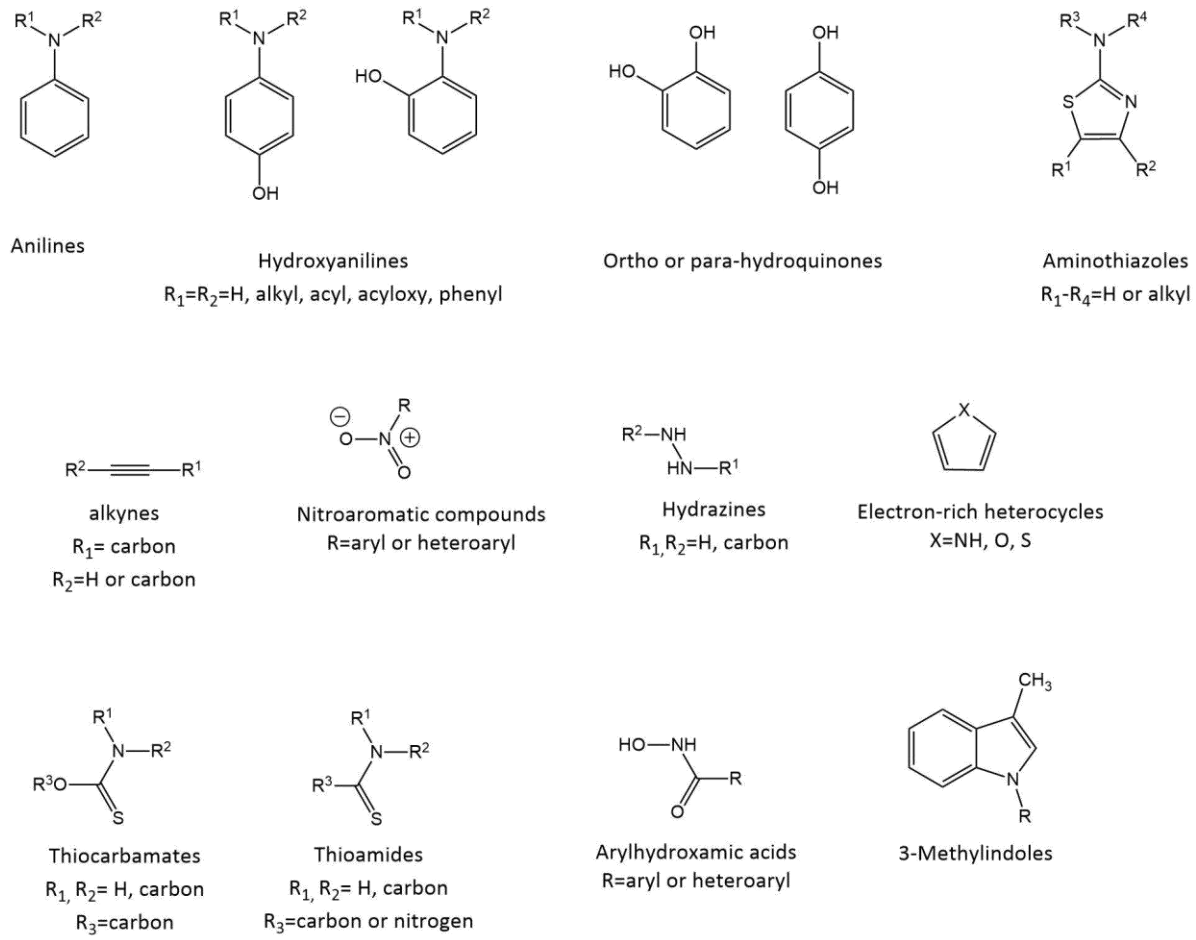


Рис. 80. Примеры структурных алертов для соединений, механизм токсичного действия которых обусловлен метаболической активацией

На Рис. 80 приведены примеры структурных алертов, указывающих на возможность органического соединения превратиться в результате метаболической активации в электрофильную молекулу, способную образовывать ковалентные связи с нуклеофильными группами биомолекул, оказывая тем самым токсичное воздействие на живой организм.

Кроме ковалентного связывания электрофилов с нуклеофильными группами макромолекул, к токсичному воздействию может приводить также ингибирование ферментов и интеркаляция в молекулы нуклеиновых кислот.

Воздействие ксенобиотиков на макромолекулы, находящихся в различных тканях живых организмов, приводят к разным видам токсичности. Так, взаимодействие с молекулами ДНК приводит к *генотоксичности* (англ. genotoxicity), под которой подразумевается *мутагенность* (англ. mutagenicity) (способность вызывать мутации в генах) и тесно связанная с ней *генотоксичная канцерогенность* (англ. genotoxic carcinogenicity) (способность приводить к появлению злокачественных новообразований в организме). Существует также *негенотоксичная канцерогенность*, которая не связана с мутациями генов [201]. Взаимодействие ксенобиотиков с белками тканей печени приводит к *гепатотоксичности* (англ. hepatotoxicity), с белками тканей почек – к *почечной токсичности* (англ. renal toxicity), с белками репродуктивных органов – *репродуктивная токсичность* (англ. reproductive toxicity), с белками либо нуклеиновыми кислотами плода – *эмбриотоксичности* (англ. embriotoxicity), с белками тканей кожи – *сенсибилизации кожи* (англ. skin sensitization) либо даже токсичности кожи (англ. skin toxicity), с белками тканей дыхательных путей – к *респираторной сенсибилизации* (англ. respiratory sensitization) либо даже токсичности (англ. respiratory toxicity). Воздействие по разным механизмам на ткани живущих в воде организмов приводит к (*острой*) *водной токсичности* (англ. acute aquatic toxicity). Особую опасность представляет *идиосинкразическая токсичность* (англ. idiosyncratic toxicity), которая встречается редко и непредсказуемо и поэтому часто не может быть выявлена во время клинических испытаний. Для каждого из вышеперечисленных видов токсичности есть свои наборы структурных алертов, многие из которых совпадают – очень редко можно найти алерт, специфичный лишь для одного вида токсичности. В Табл. 6 приведены ссылки на некоторые наборы структурных алертов, определенных для разных видов токсичности.

Табл. 6. Наборы структурных алертов, определенных для разных видов токсичности

Вид токсичности	Авторы	Ссылка
Острая водная токсичность	J.L.Hermens	[202]
	H.J.M. Verhaar et al.	[203]
Сенсибилизация кожи	M.D. Barratt et al.	[204]
	I. Gerner et al.	[205]
	M.P. Payne et al.	[206]
	S.J. Enoch et al.	[207]
Генотоксичная канцерогенность и мутагенность	J. Kazius et al.	[208]
	R. Benigni et al.	[209]
	A.B. Bailey et al.	[210]
	J. Ashby et al.	[211]
	A. Plosnik et al.	[212]
Негенотоксичная канцерогенность	R. Benigni et al.	[209]
	R. Benigni et al.	[201]
Идиосинкразическая токсичность, связанная с образованием реакционных метаболитов	A.S. Kalgutkar et al.	[213]

5.3.2.2. Структурные алерты высокой реакционной способности и нестабильности

При виртуальном скрининге обычно отбрасывают молекулы, содержащие группы с высокой реакционной способностью, поскольку такие молекулы:

- быстро разлагаются в гидролитических условиях живого организма;
- легко реагируют с биомолекулами, обуславливая высокую токсичность таких соединений;
- нестабильны в плазме крови.

При высокопроизводительном скрининге такие соединения дают ложно положительные (англ. false positive) результаты. На Рис. 81 представлены структурные алерты, которые часто используют для выявления таких молекул.

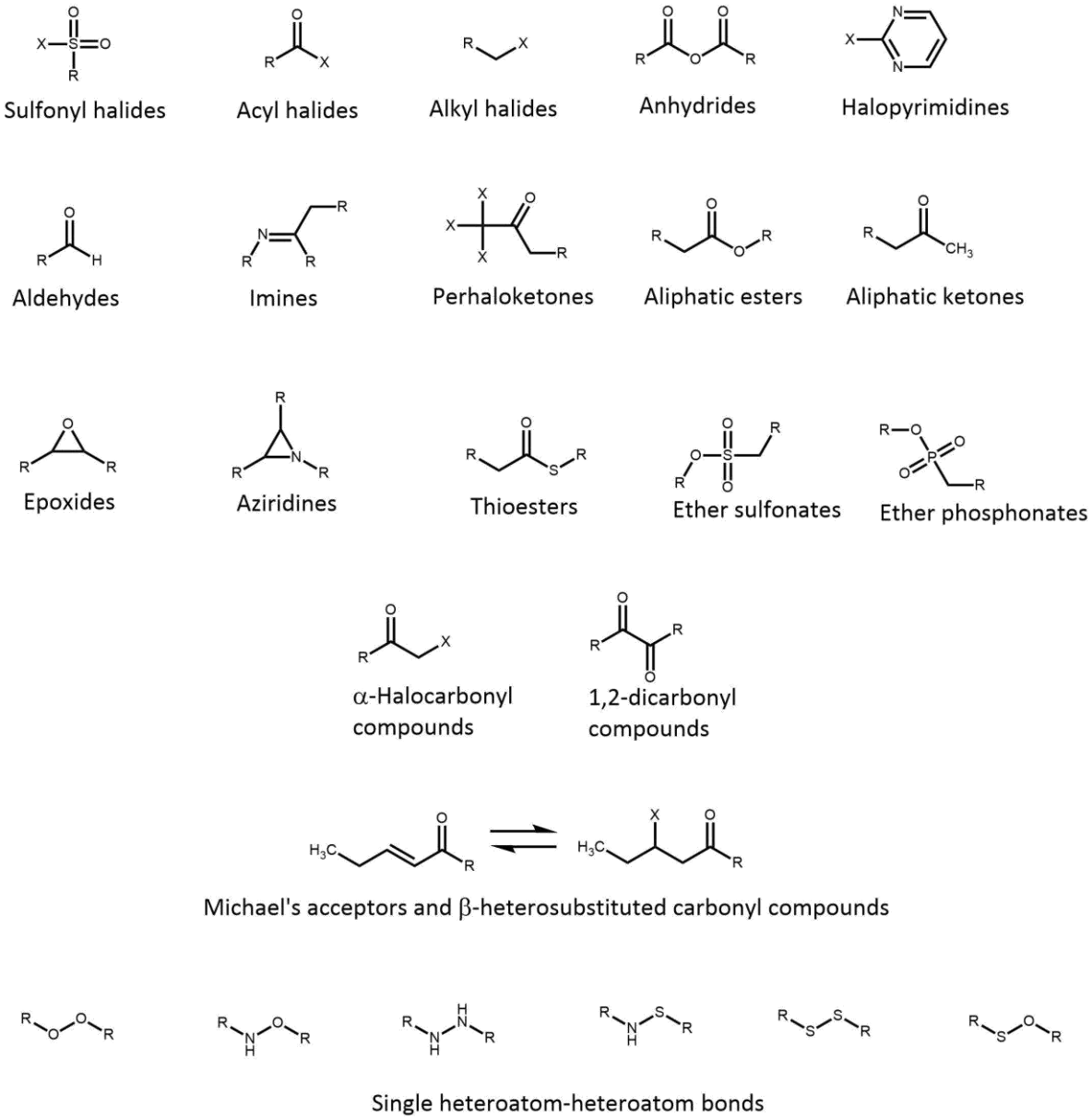


Рис. 81. Структурные алерты для соединений с высокой реакционной способностью (*X=F, Cl, Br, I,* тозил, мезил и т.д.; *R*=алкил, арил, гетероалкил, гетероарил и т.д.)

5.3.2.3. Структурные алерты PAINS

PAINS (*Pan-Assay INterference compoundS*) – это химические соединения, которые часто дают ложно положительные результаты в экспериментах по высокопроизводительному скринингу [214]. Они проявляют при этом неспецифическую активность относительно множества мишеней по множеству механизмов, включая «вмешательство» в процедуру детектирования биологической активности из-за аутофлуоресценции, образования перекиси водорода,

хелатирования металлов, агрегации молекул и других процессов [214-216]. Хотя соединения PAINS могут быть отобраны по результатам первичного скрининга как перспективные соединения-лидеры для разработки лекарственных средств, однако на последующих этапах выясняется, что они не оказывают необходимого действия на выбранные мишени. Именно поэтому таких соединений следует избегать и их отбрасывать на ранних этапах виртуального скрининга.

Для выполнения этой задачи Бэилл и Холлоуэй (Baell, Holloway) путем анализа результатов экспериментов по скринингу идентифицировали такие соединения (некоторые из них приведены на Рис. 82) и предложили набор из 480 структурных алертов PAINS (PAINS alerts), часто встречающихся в их молекулах [217]. После этого проверка на такие алерты стала необходимой частью процедуры виртуального скрининга. Возможность осуществлять такие проверки была включена в работы таких Web-сервисов, как FAF-Drug3 [218] и FAF-Drug4[1].

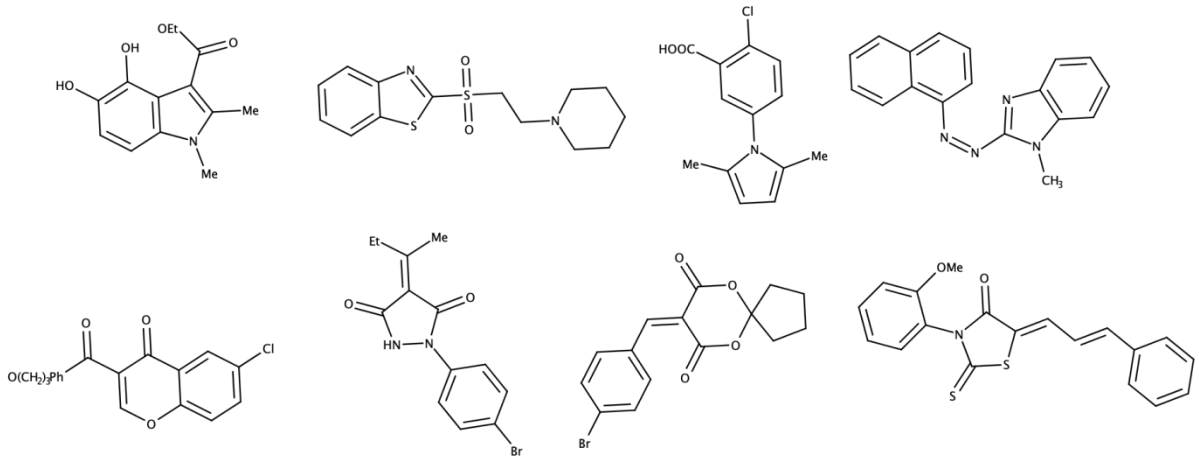


Рис. 82. Примеры соединений PAINS

Тем не менее, несмотря на популярность структурных алертов PAINS, начиная с 2017 г. практика их «слепого» использования стала подвергаться резкой критике [219, 220]. Отмечается, в частности, что эти алерты встречаются даже в одобренных лекарствах, что говорит о том, что их наличие в молекулах химических соединений должно служить лишь предупреждением, но не единственным и однозначным критерием для отбрасывания химического соединения как неперспективного для создания лекарств. Кроме того, сам набор

[1] http://fafdrugs4.mti.univ-paris-diderot.fr/pains.html

алертов в недостаточной мере оправдан с точки зрения статистики, поскольку существенная часть алертов была выведена из структур слишком малого числа соединений (часто одного) PAINS. Отмечается также необходимость учета положений соответствующих алертам PAINS подструктур в молекулах, а также необходимость учета взаимодействия между разными подструктурами в составе одной молекулы.

5.3.2.4. Принципы оценки эффективности структурных алертов

Часто используемый метод оценки эффективности структурных алертов основан на использовании *p-значений* (англ. p-values) для критерия согласия «хи-квадрат» (χ^2) Пирсона (англ. Pearson's chi-squared test). В данном случае *p*-значение показывает вероятность того, что верна «нулевая гипотеза» о том, что присутствие рассматриваемого структурного алерта не влияет на склонность химического соединения быть токсичным. Поэтому чем меньше *p*-значение, тем более статистически значимым является влияние структурного алерта на токсичность, и поэтому его можно считать более эффективным.

Табл. 7. Таблица сопряженности для влияния фрагмента на активность химического соединения

	Содержат фрагмент (*f*)	Не содержат фрагмент (*n*)	Всего
Активные (*A*)	$O_{11} = Af$	$O_{12} = An$	$N_A = Af + An$
Неактивные (*I*)	$O_{21} = If$	$O_{22} = In$	$N_I = If + In$
Всего	$N_f = Af + If$	$N_n = An + In$	$N = Af + An + If + In$

Рассмотрим процедуру вычисления *p*-значений для χ^2-критерия согласия Пирсона в применении к оценке эффективности структурных алертов. Введем следующие обозначения: *Af* – число активных (токсичных) соединений в выборке, содержащих рассматриваемый фрагмент (структурный алерт); *If* – число неактивных соединений в выборке, содержащих рассматриваемый фрагмент; *An* – число активных соединений в выборке, не содержащих рассматриваемый фрагмент; *In* – число неактивных соединений в выборке, не содержащих рассматриваемый фрагмент; *N* – общее число соединений в выборке; N_A – число активных соединений в выборке; N_I – число

неактивных соединений в выборке; N_f – число соединений в выборке, содержащих заданный фрагмент; N_n – число соединений в выборке, не содержащих заданный фрагмент. Эти значения можно поместить в *таблицу сопряженности* (англ. contingency table), представленную в Табл. 7.

Ячейки таблицы сопряженности содержат *фактические* значения числа соединений (количество наблюдений) O_{ij}. *Ожидаемое* число соединений (при условии справедливости нулевой гипотезы об отсутствии взаимосвязи между наличием фрагмента и активностью соединения) E_{ij} может быть найдено для каждой ячейки путем перемножения сумм рядов и столбцов с последующим делением на общее число соединений, см. Табл. 8.

Табл. 8. Ожидаемые значения числа соединений для таблицы сопряженности, представленной в Табл. 7

	Содержат фрагмент (f)	Не содержат фрагмент (n)	Всего
Активные (A)	$E_{11} = \frac{N_A * N_f}{N}$	$E_{12} = \frac{N_A * N_n}{N}$	N_A
Неактивные (I)	$E_{21} = \frac{N_I * N_f}{N}$	$E_{22} = \frac{N_I * N_n}{N}$	N_I
Всего	N_f	N_n	N

Значение критерия χ^2 для рассматриваемой таблицы сопряженности с двумя строками и двумя столбцами может быть вычислено по формуле:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (66)$$

Для нахождения величины p необходимо вычисленное значение сравнить с критическими значениями для распределения χ^2 с одной степенью свободы. В Табл. 9 приведены критические значения при некоторых значениях p.

Табл. 9. Таблица критических значений для распределения χ^2 с одной степенью свободы

χ^2	0.02	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
p	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Как видно из Табл. 9, чем больше значение критерия Пирсона, описывающего относительную разность между фактическим и ожидаемым (при отсутствии влияния фрагмента на активность) значением числа соединений в ячейках таблицы сопряжения, тем при меньшем *p*-значении оно достигается. Поэтому *p*-значение может случить показателем эффективности структурного алерта (фрагмента химической структуры).

Значение *p* можно также вычислить, исходя из биноминального распределения. Это значение в общем случае будет отличаться то *p*-значения для распределения χ^2. Применительно к оценке эффективности структурных алертов *p*-значение для биноминального распределения показывает вероятность того, что при отсутствии какого-либо влияния рассматриваемого фрагмента на активность химических соединений может оказаться случайным образом, что этот фрагмент присутствует в *Af* или большем числе активных соединений, когда в выборке из N соединений N_A активных. Исходя из биноминального распределения, *p*-значение может быть вычислено следующим образом [221, 222]:

$$p = \sum_{i=Af}^{N_f} \frac{N_f!}{i!\left(N_f - i\right)!}\left(\frac{N_A}{N}\right)^i\left(1 - \frac{N_A}{N}\right)^{N_f - i} \qquad (67)$$

Чем меньше это *p*-значение, тем более эффективным является структурный алерт.

Еще одной количественной характеристикой связи между факторами, задающими строки (активность соединений) и столбцы (присутствие структурного алерта) таблицы сопряженности, является *отношение шансов* (англ. odds ratio, *OR*), вычисляемое по формуле:

$$OR = \frac{Af/If}{An/In} = \frac{Af \times In}{An \times If} \qquad (68)$$

В данном случае химический смысл отношения шансов заключается в том, что оно выражает, во сколько раз доля активных соединений, содержащих данный фрагмент, больше доли активных соединений, его не содержащих.

Близкой по смыслу характеристикой является также *отношение правдоподобия* (англ. likelihood ratio), вычисляемое по формуле:

$$LR = \frac{Af/If}{N_A/N_I} = \frac{Af \times N_I}{N_A \times If} \qquad (69)$$

В данном случае смысл отношения правдоподобия заключается в том, что оно выражает, во сколько раз доля активных соединений, содержащих данный фрагмент, больше доли активных соединений по выборке.

Еще одной характеристикой, также легко вычисляемой из таблицы сопряженности, является процентная доля активных (англ. true positive rate), которая показывает, какую долю (выраженную в процентах) соединений, содержащих данный фрагмент, составляют активные:

$$TP\% = \frac{Af}{N_f} \times 100\% \qquad (70)$$

Следующая характеристика основана на теории информации Шеннона. Идея заключается в том, что «хороший» структурный алерт должен делить выборку, в которой перемешаны активные и неактивные соединения, таким образом, чтобы в подвыборке, содержащей алерт, в большей степени преобладали активные соединения, а в подвыборке, его не содержащей, в большей степени преобладали неактивные соединения. Иными словами, в результате деления выборки «хорошим» структурным алертом должна уменьшиться информационная энтропия разделения на активные и неактивные соединения. Чем более эффективным является структурный алерт, тем больше должно быть уменьшение информационной энтропии и, что эквивалентно, увеличение информации. *Увеличение информации* (англ. information gain, *IG*) является характеристикой структурного алерта, которая выражает разницу между информационной энтропией исходных данных и взвешенной (на долю соединений, содержащих или не содержащих алерт) суммой информационных энтропий подвыборок, разделенных структурным алертом:

$$IG = S\left(\frac{N_A}{N}\right) - \left(\frac{N_f}{N}\right) \times S\left(\frac{Af}{N_f}\right) - \left(\frac{N_n}{N}\right) \times S\left(\frac{An}{N_n}\right) \qquad (71)$$

где информационная энтропия Шэннона S вычисляется по формуле:

$$S(p) = -p \times \log(p) - (1-p) \times \log(1-p) \qquad (72)$$

где p – это доля активных соединений в имеющих данный структурный алерт.

Хотя все рассмотренные выше характеристики (p-значение, *OR*, *LR*, *TP*%, *IG*) выражают эффективность структурных алертов, но

делают они это по-разному, исходя из разных соображений. Очевидным недостатком *OR* и *LR* является то, что они не имеют значений (точнее, дают бесконечное значение) при отсутствии в выборке неактивных соединений, содержащих рассматриваемый алерт. В то же время использование *p*-значений является статистически обоснованным (по крайней мере, когда число соединений, содержащих алерт, превышает 10), а *IG* опирается на теорию информации. Все перечисленные характеристики активно используются при описании существующих и поиске новых структурных алертов.

5.3.2.5. Методы поиска структурных алертов

Впервые связь между структурой органических соединений и их генотоксичностью была исследована еще в конце 80-ых годов прошлого века в работе Эшби и Теннанта (Ashby и Tennant) [211], в которой был предложен первый набор структурных алертов. Такие алерты тогда формировались вручную опытными токсикологами на основании анализа опубликованных данных, причем определяющую роль при их формировании играли представления о механизме действия. Именно таким образом был сформирован основанный на алертах набор правил для разработанной в 90-ых годах экспертной системы DEREK [204, 223-225], которая до сих пор является одним из самых важных инструментов в вычислительной токсикологии[1]. Наиболее полный набор алертов генотоксичности был, сформированный таким образом Бениньи и Босса (Benigni и Bossa) [201, 209, 226], доступен через приложение с открытым кодом ToxTree[2], а также через OECD QSAR Toolbox[3].

Хотя несомненным преимуществом сформированных вручную наборов структурных алертов является опора на представления о механизмах действия, однако у такого подхода есть и существенные недостатки. Такие наборы, прежде всего, являются очень неполными, поскольку основаны на проведенном когда-то анализе имеющейся на тот момент времени литературы. Такой подход в принципе не способен охватить огромный объем публикуемого в настоящее время материала. Кроме того, далеко не все виды токсичности и далеко не всегда могут

[1] Про DEREK Nexus (от LHASA Ltd.) см. на https://www.lhasalimited.org/products/derek-nexus.htm

[2] http://toxtree.sourceforge.net/

[3] http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm

быть охарактеризованы с помощью известных представлений о механизме действия, что лишает преимущества «ручной» подход к формированию структурных алертов. Именно поэтому на первый план сейчас выходит автоматическое формирование наборов структурных алертов на основе анализа баз данных по токсичности (см. обзорные статьи [227-229]).

Можно условно выделить три основных подхода к автоматическому формированию наборов структурных алертов [229]: (1) на основе генерации фрагментов; (2) на основе «молекулярных отпечатков пальцев» (фингерпринтов); (3) на основе «интеллектуального» анализа графов (англ. graph mining).

В основе первого подхода лежит генерация всех возможных подструктур (фрагментов) определенного типа (например, всех цепочек с числом атомов, лежащих в заданных пределах). Для каждой из подструктур определяется, сколько раз она встречается в активных (токсичных) и неактивных (нетоксичных) соединениях из базы данных, и на основе этого вычисляются статистические характеристики, оценивающие эффективность его использования в качестве структурного алерта. Если подструктура чаще встречается в активных соединениях, то она называется *биофором*, а если в неактивных – *биофобом*.

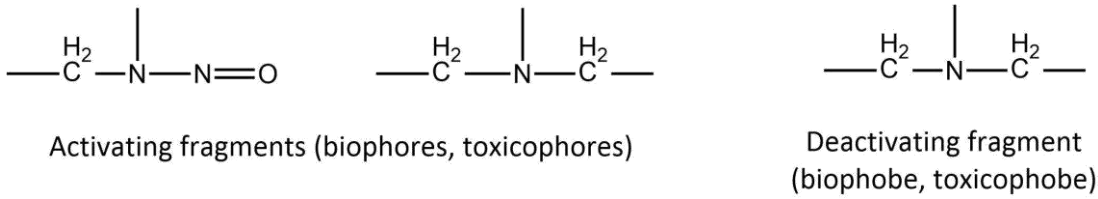


Рис. 83. Активирующие и деактивирующие фрагменты, найденные с помощью программы CASE для канцерогенности [230]

Первой программой, работающей в соответствии с этими принципами, была разработанная еще в первой половине 80-ых годов программа CASE [230, 231], основанная на генерации цепочек атомов. Эта программа сразу же была применена для поиска структурных алертов (биофоров в терминах этих работ), ответственных за мутагенное и канцерогенное действие органических соединений. На Рис. 83 приведены примеры активирующих фрагментов (биофоров, токсикофоров, структурных алертов) и деактивирующих фрагментов (биофоба, токсикофоба), выявленных с помощью программы CASE для карцерогенности.

Более развитая версия программы CASE, созданная в 90-ых годах и получившая название MULTICASE [232, 233], осуществляла иерархических анализ базы данных, когда первым идентифицировался структурный алерт с наилучшими статистическими характеристиками, вторым идентифицировался алерт с помощью анализа остатка базы данных после удаления активных соединений, содержащих первый алерт, третьим идентифицировался алерт при анализе базы данных без активных соединений, содержащий первый либо второй алерт, и т.д.

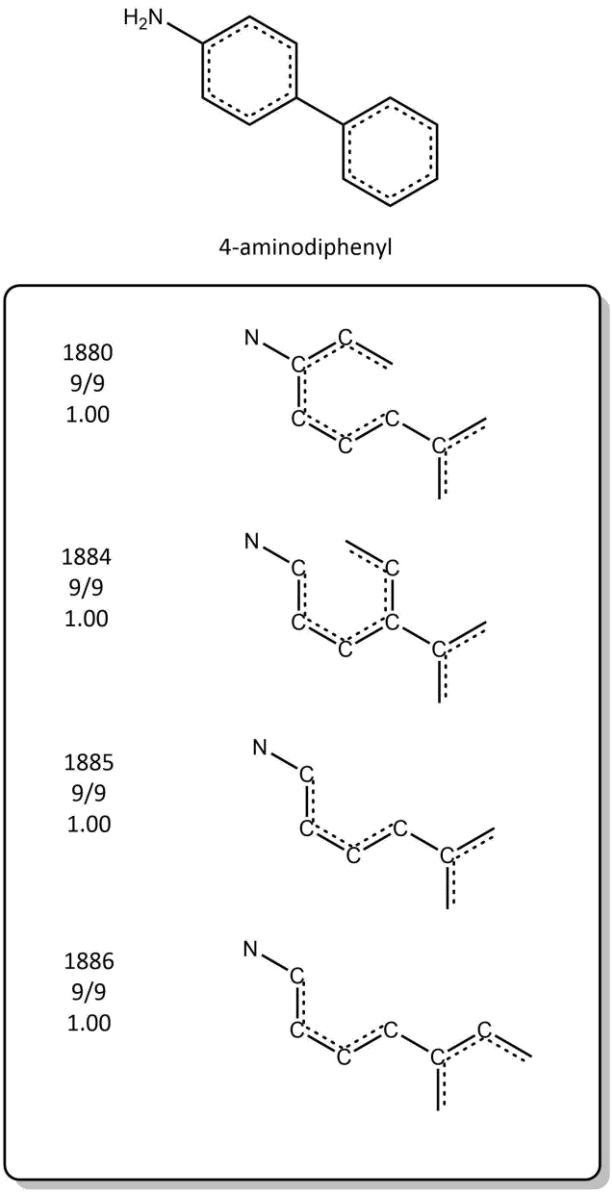


Рис. 84. 4-Аминодифенил и четыре входящие в его состав «значимые» фрагменты, используемых программой cat-SAR для прогнозирования его канцерогенной активности. Каждый из этих фрагментов содержится в 9 активных и не содержится ни в одном неактивном соединений. Рисунок адаптирован с публикации [234].

Программа cat-SAR [234] использует для идентификации структурных алертов фрагменты (линейные, циклические и разветвленные), генерируемые модулем HQSAR в составе программного комплекса SYBYL (Tripos Inc.). В рамках используемого подхода химический фрагмент считается «значимым» (англ. meaningful), если он присутствует по крайней мере в трех соединениях из обучающей выборки и если по крайней мере 90% имеющих его соединений активны (либо, наоборот, по крайней мере 90% имеющих его соединений неактивны). Вероятность обладания активностью новым соединением вычисляется исходя из того, в скольких активных и неактивных соединениях обучающей выборки присутствуют входящие в его состав «значимые» фрагменты. Например, приведенная на Рис. 84 структура 4-аминодифенила будет, согласно этому подходу, со 100% вероятностью обладать активностью, поскольку для каждого из содержащихся в нем «значимых» фрагментов из девяти содержащих его химических соединений обучающей выборки все девять активные. Программа cat-SAR была применена к анализу данных по канцерогенной активности.

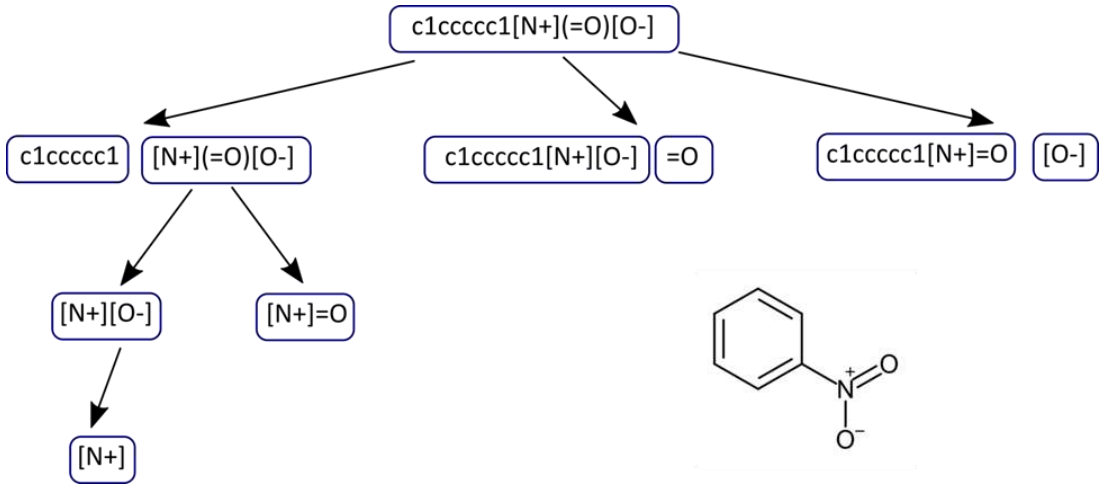


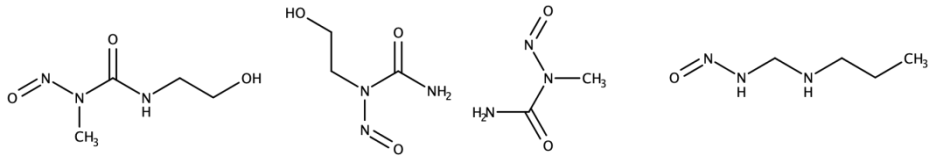
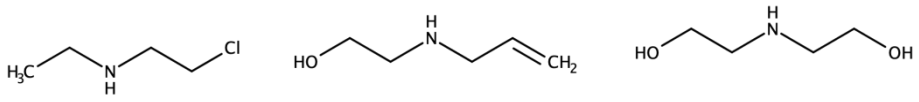
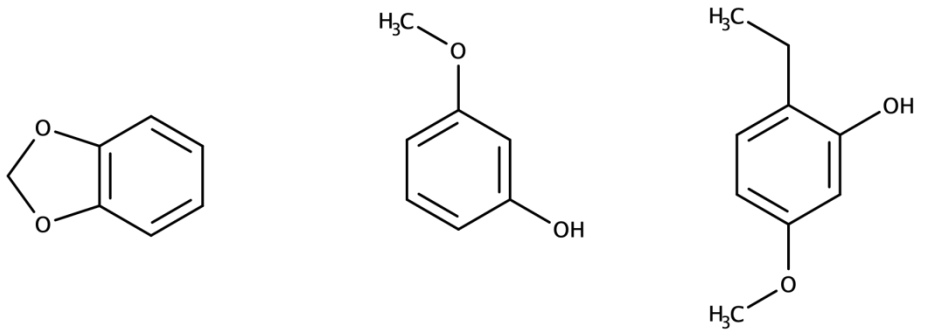
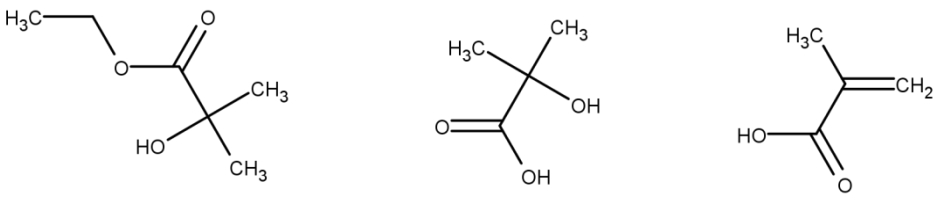
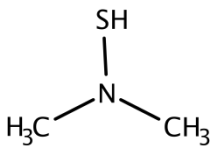
Рис. 85. Схема фрагментации строки SMILES, соответствующей молекуле нитробензола, осуществленной программой SARpy [235]

Написанная на языке Python бесплатно распространяемая программа SARpy[1] [235] основана на процедуре генерации фрагментов путем перебора различных фрагментаций представляющих химические структуры строк SMILES (см. раздел 2.2.3 в пособии 1). В качестве потенциальных структурных алертов при этом рассматриваются только целые циклы либо «ветви» структуры. На Рис.

[1] Доступна по адресу https://sourceforge.net/projects/sarpy

85 в качестве примера приведена схема фрагментации строки SMILES, соответствующей молекуле нитробензола. Эффективность структурных алертов оценивается в программе при помощи отношения правдоподобия (LR). Программа SARpy была, в частности, использована для идентификации структурных алертов канцерогенности [236]. В Табл. 10 приведены структурные формулы новых структурных алертов, идентифицированных с ее помощью.

Табл. 10. Новые структурные алерты канцерогенности, идентифицированные с помощью программы SARpy

Структурный класс	Структурные алерты
Нитрозо-мочевины	
Азотистые аналоги горчичного газа	
Бензо-диоксолы и бензодиолы	
α,β-окси- и карбокси-производные	
Замещенные серой третичные амины	

Второй подход основан на использовании стандартных наборов «молекулярных отпечатков пальцев» (фингерпринтов, структурных

ключей). В рамках данного подхода выявляются такие элементы битовой строки, которые являются ответственными за активность (то есть биты, единички в которых встречаются чаще в активных, чем в неактивных соединениях). Поскольку бит в строке соответствует определенному фрагменту, то выявление алерта не составляет труда. На использовании циркулярных фингерпринтов Моргана основана написанная на языке Python библиотека функций Bioalerts[1] для идентификации структурных алертов для токсичности [222]. В Bioalerts для оценки эффективности потенциальных структурных алертов используются *p*-значения для биноминального распределения [221]. Описанный в работе [229] метод FP основан на использовании целого набора структурных ключей, вычисляемых с помощью программы PaDEL-Descriptor[2] [237]: MACCS (166 бит) [13]; PubChem[3] (889 бит); Klekota-Roth (4860 бит)[238]; Function Group Substructure (307 бит), заданные в OpenBabel.

В третьем подходе, основанном на «интеллектуальном анализе графов» (англ. *graph mining*) химическая структура рассматривается как молекулярный граф, структурные алерты – как подграфы, и к анализу выборок применяются методы теории графов. Ключевой характеристикой в этом случае являются частоты встречаемости подграфов в анализируемых наборах графов (обучающих выборках). Алгоритмы такого рода разрабатываются главным образом математиками либо в сотрудничестве с математиками, специализирующимися в области теории графов. В частности, в методе MOLFEA [239], основанном на алгоритме Levelwise Version Space [240], выявляются линейные структурные фрагменты (цепочки атомов и соединяющих их связей, представленные в виде строк на языке SMARTS), удовлетворяющие ограничениям на минимальную и максимальную частоты встречаемости. Для выявления наиболее значимых из сгенерированных таким образом фрагментных дескрипторов проводилось построение классификационной модели при помощи методов машинного обучения. На Рис. 86 представлены 20 фрагментов, приводящих к увеличению мутагенности. Наряду со строками SMARTS здесь приведены коэффициенты, полученные в рамках метода опорных векторов (SVM) с линейным ядром. Чем более

[1] https://github.com/isidroc/bioalerts

[2] http://yapcwsoft.com/dd/padeldescriptor/

[3] ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt

положительное значение имеет этот коэффициент, тем более сильным токсикофором (структурным алертом токсичности) является структурный фрагмент.

1.627 * C:C:C:C:C:C:C:C:C
1.446 * C-Cl
1.323 * C-C-C-C-N-C
1.311 * C-C-C-O
0.952 * C-C=C
0.865 * c:c:c:c:c:n
0.824 * C-C-C-C=C
0.820 * C-C-C-N-C
0.797 * c:c:c-C=O
0.782 * C-N-C
0.780 * N-N
0.750 * C-C-C-C-O
0.723 * C-C-N-N
0.728 * N-O
0.717 * C-O-C
0.678 * C
0.674 * C-N-c:c:c:c:c:c
0.674 * C-N-c:c:c:c:c
0.672 * c:c-N
0.569 * C-C-N

Рис. 86. Идентифицированные с помощью метода MOLFEA 20 линейных фрагментов (представленных с помощью строк SMARTS) вместе с полученными с помощью метода SVM с линейным ядром коэффициентом, показывающим значимость соответствующего фрагмента [239].

Выявление структурных фрагментов более сложного строения требует применения эффективных алгоритмов работы с большим множеством графов. Такие алгоритмы, как правило, осуществляют поиск *«частых» подграфов* (англ. frequent subgraph) [241, 242] благодаря использованию свойства анти-монотонности: если какая-нибудь подструктура редко встречается в базе данных, то любая другая содержащая ее подструктура будет встречаться так же редко либо еще реже. Это позволяет организовать эффективное отбрасывание «редких» подструктур при их переборе с помощью деревьев поиска «в глубину» (англ. depth-first-search). Так, программа MoSS [243] позволяет выявлять с помощью поиска «в глубину» подструктуры, содержащие циклы и боковые цепи. Метод Gaston [244] основан на алгоритме gSpan [245], реализующий подход, основанный на «росте шаблона» (англ. Pattern-Growth Based approach) [242]. В процессе его

работы выявляются подграфы, представляющие собой циклы, пути и деревья, причем их вершины могут нести обобщенные метки, например, [N,O] для атома, который может быть азотом либо кислородом. Перечисленные выше алгоритмы были использованы для выявления структурных алертов главным образом для генотоксичности.

При анализе данных часто возникает ситуация, когда принадлежность объекта (например, химического соединения) к определенному классу (например, классу химических соединений, обладающих определенным видом активности) зависит от наличия одновременно нескольких признаков. Это нашло отражение в концепции «*возникающих образов*» (англ. emerging patterns), которые определяются как *комбинации признаков, которые значительно отличаются для разных классов объектов* [246]. Применение этой концепции к задачам хемоинформатики привело к введению понятия «*возникающих химических образов*» (англ. emerging chemical patterns) [247]. «*Прыгающие возникающие фрагменты*» (англ. Jumping Emerging Fragments, *JEF*) [248] были определены как структурные фрагменты, присутствующие в некоторых токсичных соединениях, но отсутствующие во всех нетоксичных. Для их «извлечения» их базы данных был использован алгоритм, основанный на «интеллектуальном анализе графов». Толкование признаков в «возникающих химических образах» как наличия подструктур внутри молекул привело к введению понятия «*частых возникающих молекулярных образов*» (англ. Frequent Emerging Molecular Patterns, *FEMP*), которые были использованы для поиска структурных алертов [249]. На этой основе была сформулирована концепция «*представительных обрезанных молекулярных образов*» (англ. Representative Pruned Molecular Patterns, *RPMP*) [249], в которой рассматривается одновременное наличие в молекулах активных соединений нескольких структурных фрагментов. RPMP были использованы при поиске структурных алертов. Предельным случаем RPMP, когда рассматриваемые комбинации фрагментов не присутствуют в неактивных структурах, являются «*прыгающие возникающие образы*» (англ. Jumping Emerging Patterns, *JEP*) [250]. На Рис. 87 показан пример JEP (внизу), представляющий собой комбинацию двух приведенных внизу фрагментов, а также набор активных соединений, его содержащих. Этот JET не содержится ни в одном из неактивных соединений обучающей выборки.

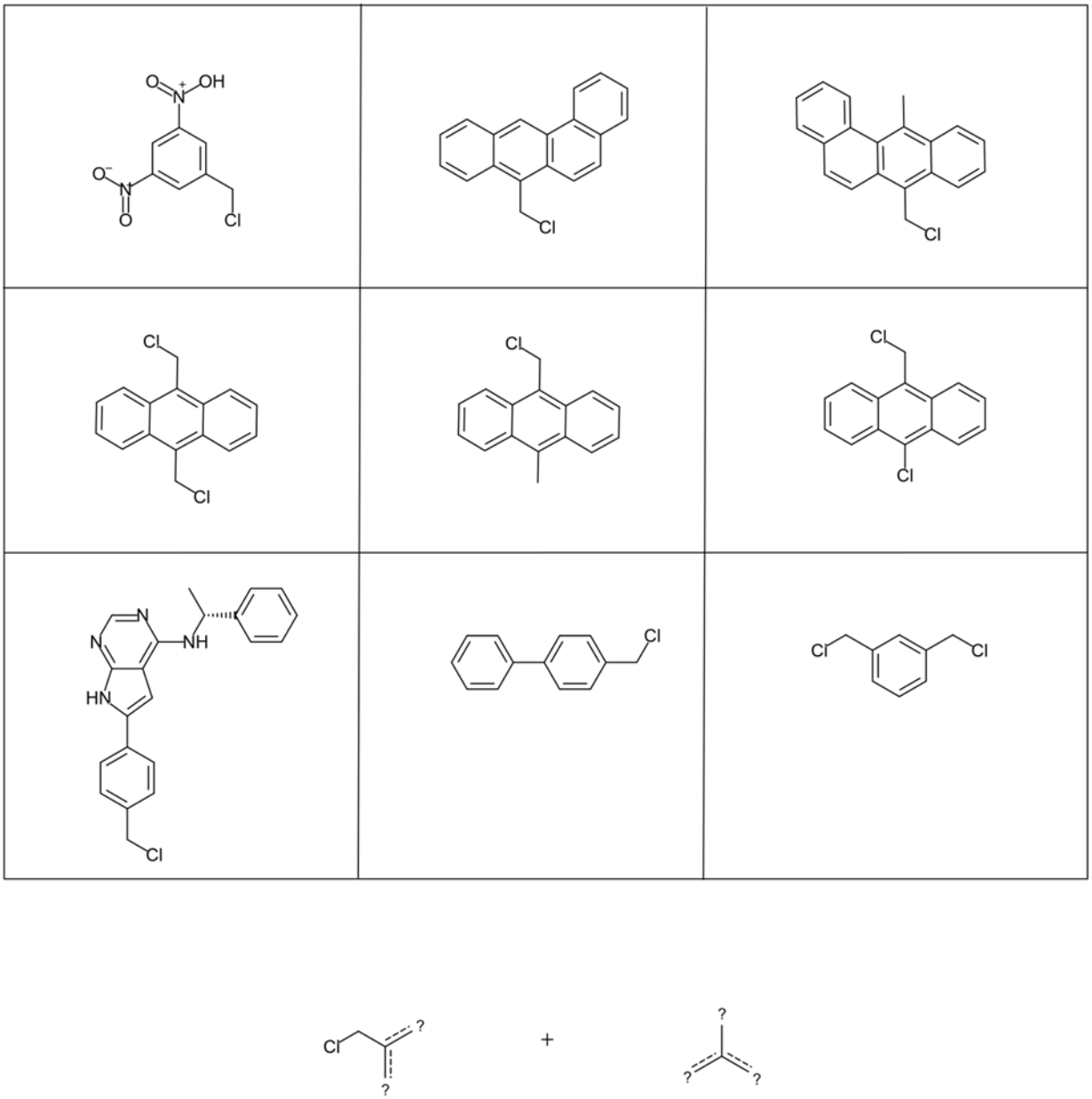


Рис. 87. Пример JEP для мутагенной активности алкилирующих соединений (внизу) и набор содержащих его активных соединений (вверху). Пример взят из публикации [250]

Наконец, для идентификации структурных алертов недавно были предложены «*стабильные возникающие молекулярные образы*» (англ. Stable Emerging Molecular Patterns, *SEMP*) [251], которые являются FEMP, обладающие наивысшими значениями специального параметра стабильности, ценность которого обосновывается в теории *формального концептуального анализа* (англ. formal concept analysis). В данном случае фрагмент, который присутствует во всех молекулах рассматриваемого множества и который нельзя расширить добавлением дополнительных атомов таким образом, чтобы

расширенный фрагмент также присутствовал во всех молекулах этого множества, называется *закрытым* относительно его. Под *стабильностью* фрагмента понимается относительное число содержащих данный фрагмент множеств молекул, относительно которых он закрыт. В цитированной публикации для «извлечения» подструктур использовался упомянутый выше алгоритм «интеллектуального анализа графов» Gaston [244].

5.3.2.6. Программные средства для проверки химических соединений на наличие структурных алертов

Хорошим программным средством для проверки химических соединений на наличие структурных алертов является Web-сервер ToxAlerts (http://ochem.eu/alerts). Этот Интернет-ресурс содержит пополняемую базу данных по разнообразным структурным алертам, взятых из различных литературных источников. Каждый алерт в базе представлен в виде строки SMARTS, и для него приведена соответствующая структура Маркуша, литературная ссылка, а также перечислены типы нежелательной активности (главным образом токсичности), на которые он ориентирован (Рис. 88). Также при помощи этого ресурса пользователь имеет возможность проверить структуры из своей базы данных на наличие структурных алертов.

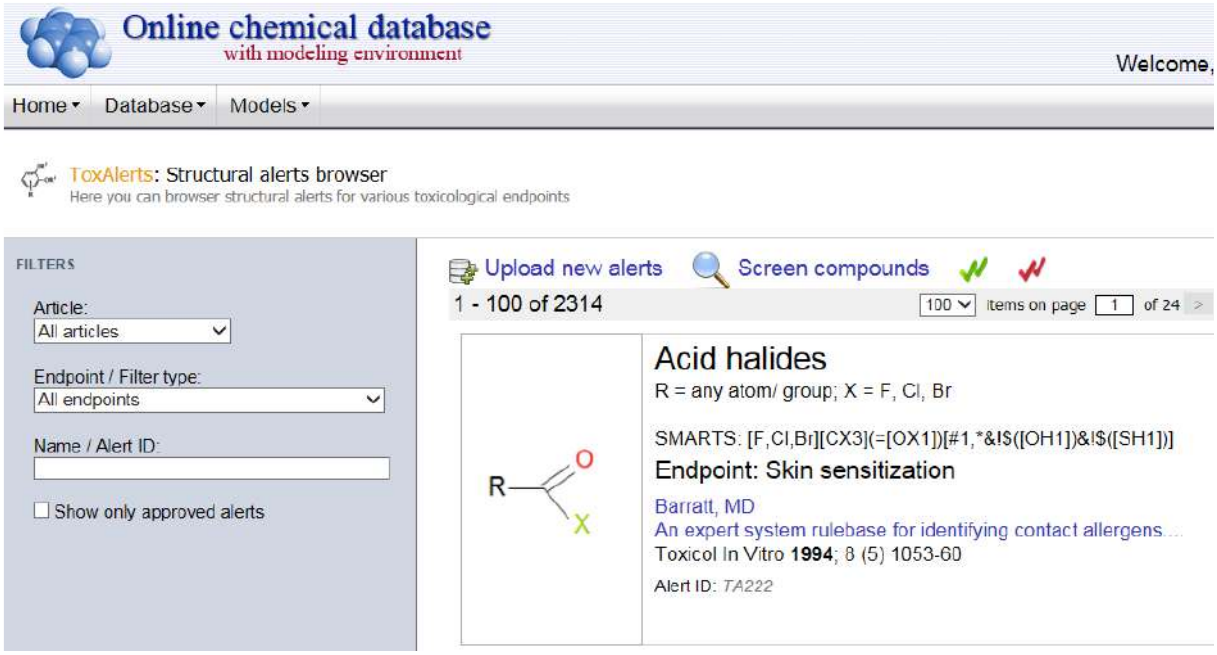


Рис. 88. Скриншот Web-сервера ToxAlerts, имплементированного на платформе OChem, http://ochem.eu/alerts

5.4. РАНЖИРОВАНИЕ МОЛЕКУЛ В ВИРТУАЛЬНОМ СКРИНИНГЕ С ИСПОЛЬЗОВАНИЕМ 2D-СТРУКТУР

При ранжировании химических соединений в виртуальном скрининге с использованием 2D-структур учитывается только связность между атомами, т.е. информация, содержащаяся в структурной формуле химического соединения. Поскольку в этом случае нет необходимости проводить анализ конформаций и строить пространственные модели, виртуальный скрининг с использованием 2D-структур осуществляется значительно быстрее по сравнению с виртуальным скринингом с помощью 3D-структур. Поэтому имеет смысл применять ранжирование с использованием 2D-структур до того, как использовать методы, основанные на рассмотрении 3D-структур молекул.

Условно можно различить четыре типа подходов к проведению виртуального скрининга с использованием 2D-структур: (1) по сходству с активными соединениями; (2) по склонности к обладанию нужным видом активности, оцениваемой при помощи одноклассовых моделей; (3) по вероятности обладания нужным видом активности, оцениваемой при помощи двухклассовых моделей; (4) по количественному значению активности, спрогнозированному при помощи регрессионной модели «структура-свойство»

5.4.1. Ранжирование по сходству с активными соединениями

Возможность проведения виртуального скрининга по сходству с активными соединениями связана с предположением о том, что сходные структуры имеют сходные свойства. В этом случае в качестве оценки (скоринга) для ранжирования химических соединений выступает мера сходства с одним или несколькими специально выбранными химическими соединениями, обладающими желаемым видом активности. Хотя любая из указанных в разделе 1.2.2 мер сходства может быть использована для этой цели, наибольшей популярностью пользуется индекс Танимото в сочетании с представлением молекул с помощью одного из видов молекулярных «отпечатков пальцев» (фингерпринтов). В случае, если рассматривается только одно активное соединение, то виртуальный скрининг сводится к поиску по сходству в базе данных с использованием в этой структуры в качестве запроса. В статье [252] проведено сравнение мер сходства для векторов небинарных дескрипторов.

5.4.1.1. Скрининг с использованием нескольких активных структур. Понятие о «слиянии данных» (data fusion)

В случае, если рассматривается несколько активных структур, принято комбинировать вычисленные по отношению к ним значения мер сходства при помощи процедуры, получившей название «*слияния данных*» (англ. data fusion), см. обзор [253]. Существует две основных разновидности этой процедуры – «без учителя» (англ. *unsupervised*) и «с учителем» (англ. *supervised*).

«Слияние данных» «без учителя» является наиболее часто используемым. Оно сводится к формальному комбинированию мер сходства по отношению к n (n>1) активным соединениям и используется главным образом на начальных этапах поиска новых лекарственных препаратов, когда объем данных «структура-активность» очень ограничен. В Табл. 11 приведены наиболее часто используемые правила комбинирования мер сходства S_{ij}, вычисленных по отношению к i-ому активному соединению для j-ого соединения из скринируемой базы данных либо их рангов (R_{ij}). Например, правило MAX говорит о том, что для ранжирования скринируемой базы данных в качестве комбинированной меры сходства для соединения j используется максимальное из значений сходства, вычисленных по отношению ко всем рассматриваемым активным соединениям i. В правилах ANZ и MNZ рассматриваются не все n активных соединений, а только те p ($p \leq n$) из них, мера сходства с которыми превышает определенный порог. В основанном на рассмотрении рангов R_{ij} правиле RRF рассматриваются p активных соединений, входящих в верхний 1% ранжированного списка.

Процедура «с учителем» сводится к построению классификационной модели, опирающуюся, как правило, на теорию вероятностей, с использованием в качестве дескрипторов мер сходства, вычисленным по отношению к активным соединениям. Очевидно, для построения таких моделей требуется наличие достаточно большого объема экспериментальных данных «структура-активность».

Табл. 11. Правила комбинирования мер сходства или их рангов в рамках процедуры «слияния данных» «без учителя»

Название	Формула
MAX	$max\{S_{1j}, S_{2j}, \dots, S_{ij}, \dots, S_{nj}\}$
MIN	$min\{S_{1j}, S_{2j}, \dots, S_{ij}, \dots, S_{nj}\}$
SUM	$\frac{1}{n}\sum_{i=1}^{n} S_{ij}$
MED	$median\{S_{1j}, S_{2j}, \dots, S_{ij}, \dots, S_{nj}\}$
ANZ	$\frac{1}{p}\sum_{i=1}^{p} S_{ij}$
MNZ	$p\sum_{i=1}^{p} S_{ij}$
EUC	$\sqrt{\sum_{i=1}^{n} S_{ij}^2}$
RRF	$\sum_{i=1}^{p} \frac{1}{R_{ij}}$

Рассмотрим в качестве примера подход, предложенный в публикации [254], в основе которого лежит преобразование меры сходства S_x по отношению к какому-либо из выбранных активных соединений в значение вероятности активности $P(A|S_x)$ по формуле Байеса:

$$P(A|S_x) = \frac{P(S_x|A)P(A)}{P(S_x)} \propto \frac{P(S_x|A)}{P(S_x)} \qquad (73)$$

где: $P(A|S_x)$ – условная вероятность того, что соединение со значением меры сходства S_x по отношению к выбранной активной структуре также является активным; $P(S_x|A)$ – условная вероятность того, что между выбранной активной структурой и произвольной активной структурой из базы данных будет мера сходства S_x; $P(S_x)$ – безусловная вероятность получения меры сходства S_x; $P(A)$ – безусловная вероятность наличия активности у произвольной молекулы из базы данных. Поскольку $P(A)$ не зависит от меры сходства и является поэтому константой, этот член может быть выведен из рассмотрения.

Оценку $P(S_x)$ можно получить путем анализа распределения значений мер сходства в базе данных, а $P(S_x|A)$ – путем анализа распределения значений мер сходства внутри кластеров, образованных активными соединениями. Далее делается предположение, что условные вероятности $P(A|S_x)$ по отношению к разным активным соединениям описывают независимые события, и поэтому комбинированное значение вероятности получается путем перемножения этих вероятностей. Далее комбинированные значения вероятности могут быть использованы для ранжирования скринируемой базы данных для целей виртуального скрининга.

В работе [255] было предложено проводить «слияние данных» путем построения модели при помощи логистической регрессии:

$$FS_j = \frac{1}{1 + e^{-(a+\sum_i b_i S_{ij})}} \qquad (74)$$

где FS_j – комбинированная мера сходства для j-ого соединения из скринируемой базы данных, a и b_i (i=1,…,n) – настраиваемые коэффициенты, получаемые путем обработки обучающей выборки, содержащей активные и неактивные соединения.

С другими методами «слияния данных» «с учителем» можно ознакомиться по обзорной статье [253].

5.4.2. Ранжирование по склонности к обладанию нужным видом активности, оцениваемой при помощи одноклассовых моделей

Понятие одноклассовой классификации рассмотрено в разделе 2.14.4 Пособия 4. Методы одноклассовой классификации позволяют отличать объекты, принадлежащие целевому (англ. target) классу от объектов, этому классу не принадлежащих. Методы одноклассовой классификации, кроме прогнозирования принадлежности целевому классу еще, как правило, выдают характеристику, показывающую степень уверенности в прогнозе. Такие характеристики показывают степень близости анализируемого объекта целевому классу, использованному для построения модели. Если целевой класс состоит из химических соединений, обладающих биологической активностью определенного типа, то с помощью такой характеристики можно ранжировать химические соединения по склонности к обладанию таким же видом биологической активности. Это позволяет использовать такие характеристики в качестве оценки (скоринга) для виртуального скрининга.

Практически все методы одноклассовой классификации могут быть использованы для проведения виртуального скрининга.

Например, в случае одноклассового метода опорных векторов (1-SVM) (см. раздел 2.5.5 в Пособии 4) в качестве оценки (скоринга) выступает расстояние в пространстве признаков между тестируемым объектом и разделяющей гиперплоскостью – чем дальше от гиперплоскости в сторону, противоположную началу координат, находится тестируемая молекула, тем более вероятно у нее наличие того типа биологической активности, для которого одноклассовая модель была построена. Возможность применять 1-SVM для проведения виртуального скрининга была показана в работе [256] для случая фрагментных дескрипторов и в работах [257, 258] при описании молекул при помощи непрерывных молекулярных полей. При построении одноклассовых моделей при помощи автокодировщиков на основе многослойных нейронных сетей (см. раздел 2.11.3.3 в Пособии 4) в качестве скоринга выступает ошибка восстановления данных. Эффективность применения этой характеристики при проведении виртуального скрининга была продемонстрирована в публикации [259]. При построении одноклассовых моделей при помощи ограниченной машины Больцмана (RBM) (см. раздел 2.11.7.3 в Пособии 4) в качестве скоринга для проведения виртуального скрининга могут быть использованы как ошибка восстановления данных, так и свободная энергия, что было продемонстрировано в публикации [260].

5.4.3. Ранжирование по вероятности обладания нужным видом активности, оцениваемой при помощи классификационных моделей

Виртуальный скрининг может быть осуществлен при помощи ранжирования по вероятности обладания нужным видом активности, оцениваемой при помощи классификационных моделей. Для этой цели может быть применен любой классификационный метод машинного обучения, позволяющий при прогнозировании давать вероятностную оценку уверенности в принятом решении. Для этой цели чаще всего используется наивный байесовский классификатор и его модификации (см. раздел 2.6.2 в Пособии 4) [261], метод случайного леса (см. раздел 2.8.2 в Пособии 4) [262], метод ближайших соседей (см. раздел 2.4.1 в Пособии 4) [263], искусственные нейронные сети (см. раздел 2.11 в Пособии 4) [264]. Из методов построения классификационных моделей, которые при прогнозе не дают вероятностную оценку, наибольшей популярностью при проведении виртуального скрининга пользуется метод опорных векторов SVM (см. раздел 2.5 в Пособии 4) [265]. В последнем случае в качестве скоринга выступает расстояние

между тестируемым объектом и гиперплоскостью в пространстве признаков. Сравнительный анализ возможности использования различных классификационных методов в виртуальном скрининге представлен в работе [266].

5.4.4. Ранжирование по количественному значению активности, спрогнозированному при помощи регрессионной модели «структура-свойство»

При проведении виртуального скрининга химические соединения могут быть ранжированы либо отобраны в соответствии со значениями, спрогнозированными при помощи регрессионной модели «структура-свойство» (QSAR/QSPR). В этом случае обычно рассматриваются только те соединения, которые попадают в область применимости этой модели (см. раздел 2.4 в Пособии 3). При использовании виртуального скрининга для поиска новых лекарственных препаратов такой вид ранжирования используют лишь на заключительных стадиях исследования, когда накоплен достаточный объем экспериментальных данных для построения моделей QSAR. Напротив, при разработке новых материалов отбор при помощи моделей QSPR может составлять основу всего цикла исследований.

5.5. РАНЖИРОВАНИЕ ХИМИЧЕСКИХ СОЕДИНЕНИЙ В ВИРТУАЛЬНОМ СКРИНИНГЕ С ИСПОЛЬЗОВАНИЕМ 3D-СХОДСТВА С АКТИВНЫМИ СТРУКТУРАМИ

Биологическая активность химических соединений часто бывает обусловлена специфическим связыванием с присутствующими в живых организмах биологическими макромолекулами, такими как ферменты, рецепторы, нуклеиновые кислоты. При разработке новых лекарств о таких макромолекулах и о сайтах связывания внутри них говорят как о мишенях действия молекул лекарств. Отличительной особенностью биологических макромолекул, являющихся мишенями действия молекул лекарств, является четко выраженное и зачастую фиксированное пространственное строение. Поскольку сильное связывание молекул лиганд с биологическими макромолекулами предполагает комплементарность их пространственных форм (по крайней мере, комплементарность пространственной формы молекулы низкомолекулярного лиганда и пространственной формы сайта связывания внутри макромолекулы), то пространственные формы

молекул, специфически связывающихся с одним и тем же сайтом на той же самой макромолекуле и поэтому обладающих сходным типом биологической активности, должны быть сходными между собой. Вследствие этого ранжирование молекул в виртуальном скрининге с использованием сходства их пространственного строения (3D-сходства) с пространственным строением молекул биологически активных молекул является эффективным подходом к дизайну новых биологически активных соединений.

Существует несколько подходов к оценке 3D-сходства молекул. В частности, у обладающих 3D-сходством молекул должны быть сходны функции электронной плотности, их молекулярные формы, а также физические поля, с помощью которых они взаимодействуют с биологическими мишенями. Сравнение пространственного строения молекул может как включать, так и не включать их совмещение в пространстве. При установлении 3D-сходства для гибких молекул может рассматриваться одна либо множество фиксированных конформаций для каждой молекулы, либо сама процедура сравнения может предполагать рассмотрение гибкости их молекул. Кроме того, может рассматриваться как попарное 3D-сходство молекул, так и 3D-сходство рассматриваемой молекулы сразу с набором биологически-активных молекул (т.н. консенсусное сходство). Все это приводит к большому разнообразию методов оценки 3D-сходства молекул.

5.5.1. Квантовое сходство. Индекс Карбо

Исторически первым подходом к количественной оценке 3D-сходства молекул было предложенное Р. Карбо «квантовое сходство» [267], вычисляемое как нормированный интеграл произведения функций электронной плотности $\rho_A(\mathrm{r})$ и $\rho_B(\mathrm{r})$ двух молекул A и B (т.н. индекс Карбо):

$$S_{A,B} = \frac{\iiint \rho_A(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r}}{\sqrt{\iiint \rho_A(\mathbf{r})\rho_A(\mathbf{r})d\mathbf{r}} \cdot \sqrt{\iiint \rho_B(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r}}} \qquad (75)$$

Максимальное значение индекса Карбо, равное единице, достигается, когда функции электронной плотности совпадают, т.е. $\rho_A(\mathrm{r})=\rho_B(\mathrm{r})$ во всех точках пространства. Полная идентичность функций электронной плотности двух молекул возможна только тогда, когда они химически идентичны, находятся в одной и той же конформации (если молекулы гибкие) и одинаковым образом расположены и ориентированы в пространстве (т.е. совмещены). Во всех остальных случаях значение

$S_{A,B}$ меньше единицы, но больше нуля. Чем больше значение $S_{A,B}$, тем выше их 3D-сходство.

Поскольку значение квантового сходства зависит от взаимного расположения и ориентации двух молекул, при его вычислении проводят их совмещение (выравнивание, англ. *alignment*) в пространстве. В этом случае одну из них используют в качестве неподвижного шаблона, а вторую ориентируют относительно первой таким образом, чтобы значение индекса Карбо было наибольшим:

$$S_{A,B} \to max \leftrightarrow \iiint \rho_A(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r} \to max \qquad (76)$$

Тем самым концепция квантового сходства естественным образом определяет подход к совмещению молекул в пространстве.

Хотя способ оценки 3D-сходства молекул путем сравнения функций их электронной плотности является логичным и вполне естественным, однако практическое его применение часто сталкивается с необходимостью проводить квантово-химические расчеты большой массы молекул, а также осуществлять трудоемкую процедуру их попарного совмещения. Вследствие этого при виртуальном скрининге больших баз данных принцип квантового сходства вряд ли может быть использован. Тем не менее, этот способ оценки 3D-сходства молекул имеет большое историческое значение, поскольку многие из современных подходов все равно отчасти базируются на связанных с ним идеях.

5.5.2. Сходство пространственных форм молекул

5.5.2.1. Понятие о пространственной форме молекулы

В настоящее время большинство подходов к оценке 3D-сходства молекул основано на концепции *пространственной формы молекулы* (англ. molecular shape). Существование формы у молекул обусловлено формой кривой потенциала взаимодействия между несвязанными атомами, которую можно разделить на две части, одна из которых ответственна за слабое притяжение на больших расстояниях благодаря дисперсионному взаимодействию, а другая – за сильное отталкивание на малых расстояниях из-за невыгодного обменного взаимодействия (см. Рис. 89). Если первая из этих частей является относительно пологой и в эмпирических силовых полях описывается как зависимость, пропорциональная r^{-6}, то вторая часть описывает резкое возрастание потенциала по мере сближения ядер атомов, и эту крутую

зависимость обычно аппроксимируют как r^{-12} либо как обратную экспоненту.

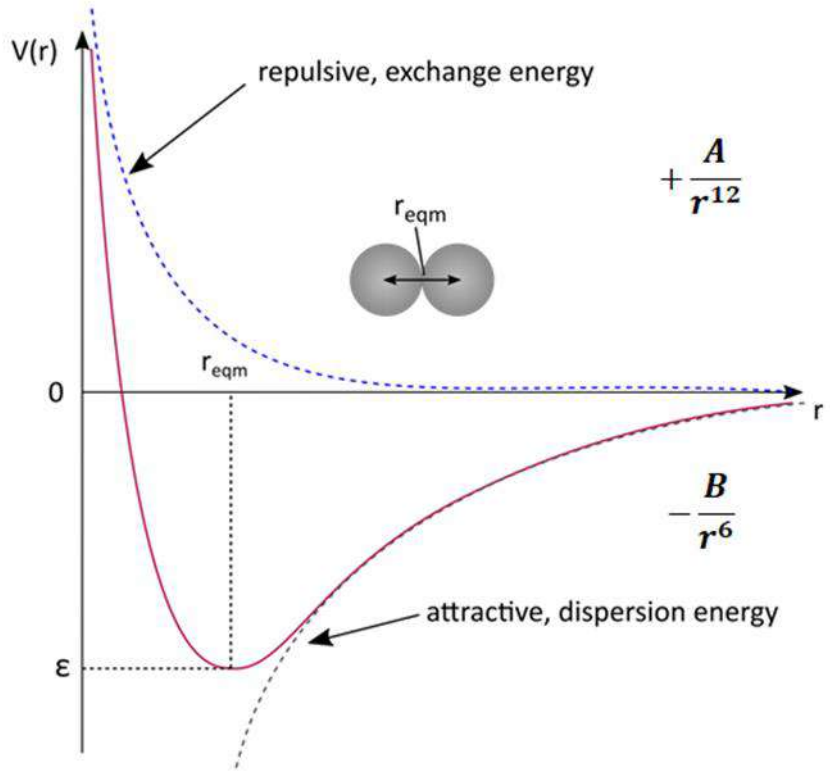


Рис. 89. Потенциал взаимодействия $V(r)$ между двумя несвязанными атомами как функция межъядерного расстояния r.

Прямым следствием крутизны отталкивающей части потенциала межатомного взаимодействия является невозможность атомов «проникать» друг в друга. Вследствие этого область пространства, содержащая атомы одной молекулы, является недоступной для атомов других молекул. Это естественным образом приводит к понятию молекулярной формы. Таким образом, *пространственная форма молекулы – эта область пространства, занимаемая атомами одной молекулы, которая недоступна для атомов других молекул*.

5.5.2.2. Принцип комплементарности пространственных форм молекул

При образовании прочных комплексов между двумя молекулами белка, между молекулой белка и низкомолекулярного лиганда (например, молекулой лекарства), а также при упаковке молекул внутри кристаллической решетки, их молекулярные формы оказываются *комплементарными* друг другу, когда выпуклые области на поверхности одной из них находятся в контакте с вогнутыми областями другой молекулы (Рис. 90).

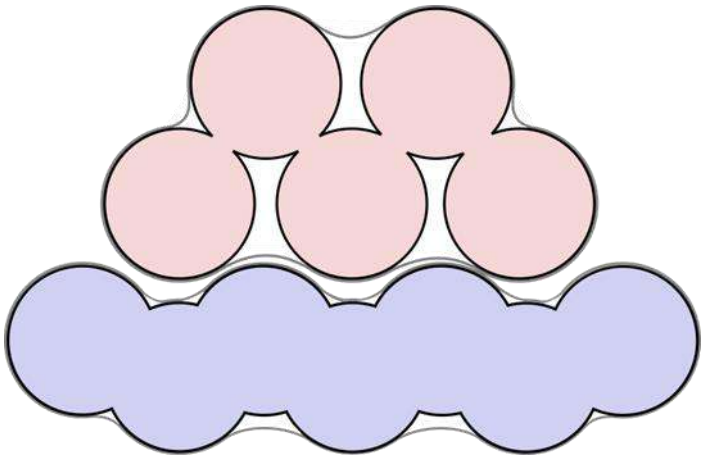


Рис. 90. Принцип комплементарности форм: выпуклые области одной из них находятся в контакте с вогнутыми областями другой формы.

Комплементарность форм молекул важно для образования прочных межмолекулярных комплексов по следующим причинам. Во-первых, при комплементарности форм максимизируется энергия ван-дер-ваальсова взаимодействия между молекулами, что приводит к большому выигрышу в энтальпии взаимодействия. Во-вторых, благодаря комплементарности форм максимальное число молекул воды уходит в основной объем растворителя, где они обладают большим числом степеней свободы, что приводит к большому выигрышу в энтропии.

Благодаря свойству комплементарности, форма молекулы лиганда оказывается комплементарной форме сайта связывания в макромолекуле биологической мишени (см. Рис. 91) и, следовательно, лиганды, обладающие близкими формами имеют тенденцию быть комплементарными и, следовательно, образовывать прочные комплексы с одним и тем же сайтом связывания одной и той же биологической мишени, т.е. обладать одним и тем же видом биологической активности. Именно поэтому можно ожидать, что новые молекулы лекарств могут быть найдены путем поиска в процессе виртуального скрининга либо генерации молекул, обладающих 3D-сходством с активными молекулами.

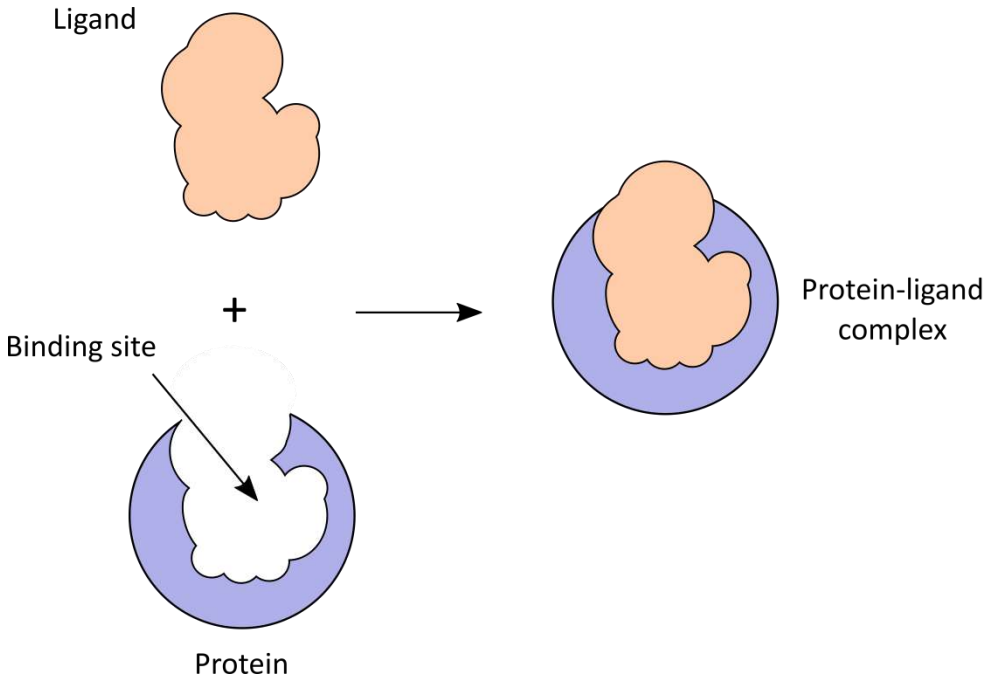


Рис. 91. Комплементарность формы лиганда форме сайта связывания биологической мишени

5.5.2.3. Сравнение пространственных форм молекул в модели жестких сфер

Количественные меры оценки сходства пространственных форм молекул *A* и *B* основаны на вычислении их объемов V_A и V_B, а также объемов их пересечения $V_{A\cap B}$ и объединения $V_{A\cup B}$ (Рис. 92). Под пересечением пространственных форм двух молекул понимается множество точек, принадлежащих одновременно пространственным формам обеих молекул, тогда как под объединением понимается множество точек, присутствующих хотя бы в одной молекуле. Вычисление объемов пересечения и объединения пространственных форм требует совмещения молекул в пространстве.

Для количественной оценки сходства пространственных форм молекул чаще всего используют два типа характеристик – расстояние между формами (англ. *shape distance*) и индекс Танимото для форм (англ. *Tanimoto shape index*), которые вычисляются по формулам:

$$D_{A,B} = \sqrt{V_A + V_B - 2V_{A\cap B}} \quad (77)$$

$$T_{A,B} = \frac{V_{A\cap B}}{V_{A\cup B}} = \frac{V_{A\cap B}}{V_A + V_B - V_{A\cap B}} \quad (78)$$

Преимуществом использования расстояния является то, что эта характеристика является метрикой, удовлетворяющей правилу треугольника. Нулевое расстояние соответствует полной

идентичности пространственных форм, тогда как максимальное значение достигается при отсутствии пересечения в пространстве между молекулами и зависит от объемов этих молекул. В то же время индекс Танимото имеет четкие границы. Максимальное значение индекса Танимото равно единице и достигается при идентичности пространственных форм, тогда как минимальное значение равно нулю и достигается тогда, когда молекулы не пересекаются в пространстве.

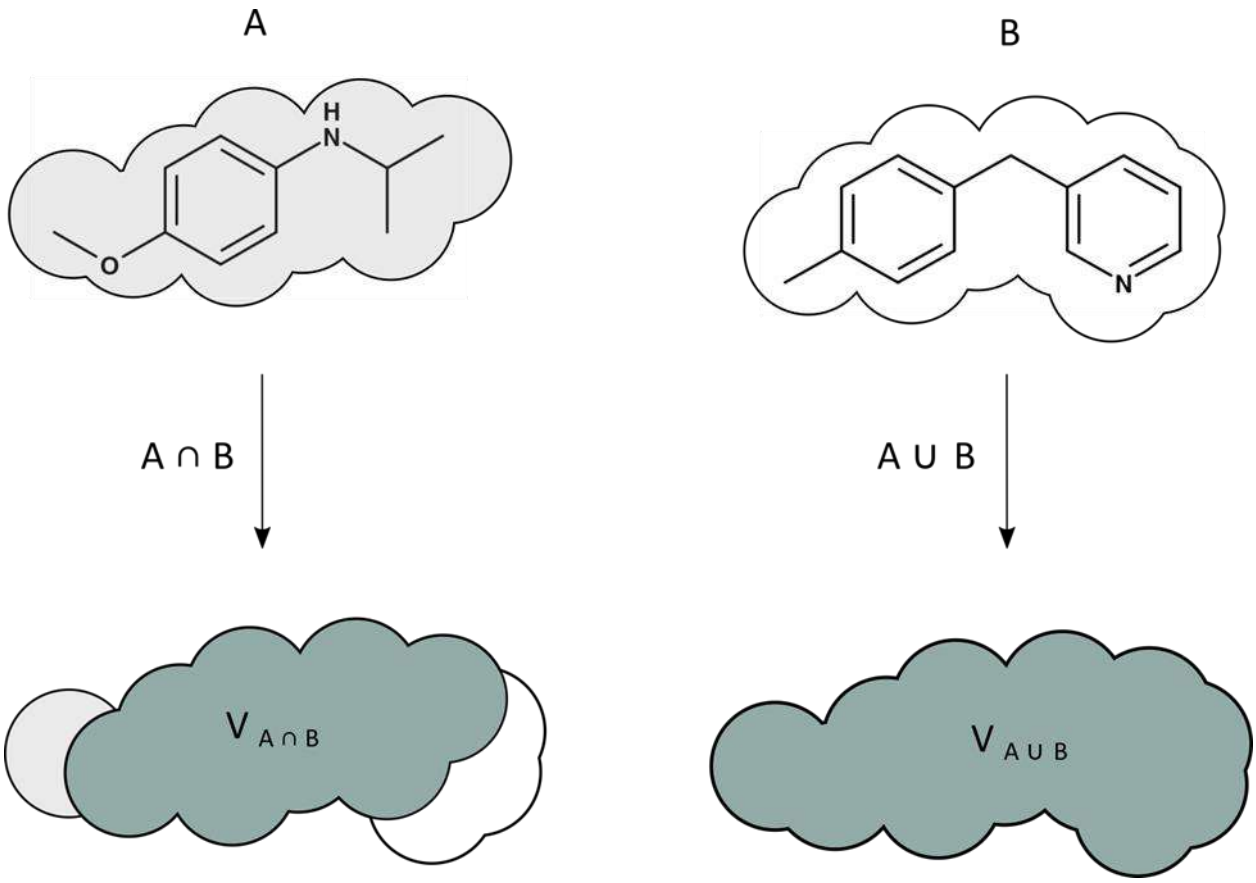


Рис. 92. Пересечение и объединение пространственных форм молекул

Концептуально наиболее естественным способом вычисления объемов пространственных форм молекул, их пересечения и объединения, необходимых для количественной оценки их сходства, является применение модели жестких сфер, согласно которой каждый атом моделируется при помощи сферы с центром на ядре атома и с радиусом, равным его ван-дер-ваальсову радиусу. В этом случае объем атома может быть вычислен при помощи хорошо известной из геометрии формулы объема сферы (шара):

$$V = \frac{4}{3}\pi R^3 \qquad (79)$$

Найти объем пересечения двух сфер уже значительно сложнее, но все еще возможно в аналитическом виде[1]:

$$V_{a\cap b} = \frac{\pi\left(R_a + R_b - d_{a,b}\right)^2\left(d_{a,b}^2 + 2d_{a,b}(R_a + R_b) - 3(R_a^2 + R_b^2) + 6R_aR_b\right)}{12d_{a,b}} \qquad (80)$$

где: R_a – радиус атома a; R_b – радиус атома b; $d_{a,b}$ – расстояние между ядрами атомов a и b. Отметим, что в этой формуле использованы строчные буквы a и b для обозначения атомов, в отличие от прописных букв A и B, примененных в формулах выше для обозначения молекул. На Рис. 93 представлен случай пересечения трех сфер. В этом случае нужно вычислить не только их объемы, но их объемы их трех попарных пересечений, а также объем их тройного пересечения $V_{a\cap b\cap c}$:

$$V = V_a + V_b + V_c - V_{a\cap b} - V_{a\cap c} - V_{b\cap c} + V_{a\cap b\cap c} \qquad (81)$$

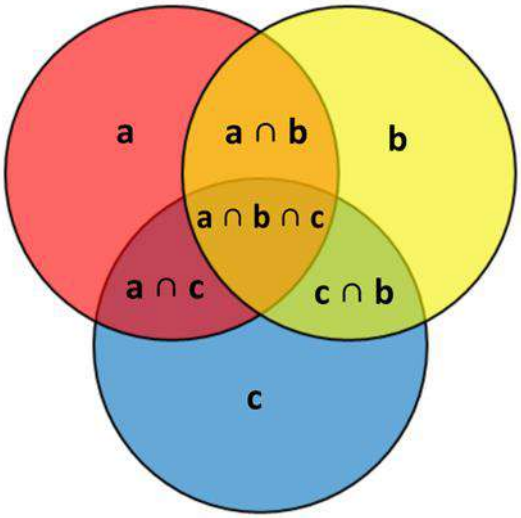


Рис. 93. Пересечение трех сфер

Нахождение объема тройного пересечения сфер является еще более сложной задачей и уже с трудом может быть решена в аналитическом виде. При еще большем числе сфер, как это имеет место в случае реальных молекул, состоящих из множества атомов, задача становится практически неразрешимой в аналитическом виде. Тем не менее, может быть получено численное решение задачи путем численного интегрирования с использованием густой решетки точек либо по методу Монте-Карло.

[1] Вывод этой формулы представлен на http://mathworld.wolfram.com/Sphere-SphereIntersection.html

Запишем для этого выражение для объема молекулы, состоящей из N атомов, индексируя атомы их порядковыми номерами в молекуле:

$$V = \sum_{1 \le i \le N} V_i - \sum_{1 \le i < j \le N} V_{i,j} + \sum_{1 \le i < j < k \le N} V_{i,j,k} - \ldots \quad (82)$$

где V_i – объем атома i, $V_{i,j}$ – объем пересечения атомов i и j, $V_{i,j,k}$ – объем пересечения трех атомов i, j и k и т.д. Обозначим через $\delta_i(\mathbf{r})$ индикаторную функцию принадлежности точки с радиус-вектором $\mathbf{r}$ атому i:

$$\delta_i(\mathbf{r}) = \begin{cases} 1, & |\mathbf{r} - \mathbf{R}_i| \le R_i \\ 0, & |\mathbf{r} - \mathbf{R}_i| > R_i \end{cases} \quad (83)$$

где $\mathbf{R}_i$ – радиус-вектор положения ядра атома i в пространстве, R_i – ван-дер-ваальсовый радиус атома i. В этом случае индикаторная функция принадлежности точки всей молекуле, которая будет обозначена как $\delta(\mathbf{r})$, может быть выражена следующим образом:

$$\delta(\mathbf{r}) = \sum_{1 \le i \le N} \delta_i(\mathbf{r}) - \sum_{1 \le i < j \le N} \delta_i(\mathbf{r})\delta_j(\mathbf{r}) + \sum_{1 \le i < j < k \le N} \delta_i(\mathbf{r})\delta_j(\mathbf{r})\delta_k(\mathbf{r}) - \ldots \quad (84)$$

Можно показать, что это выражение может быть представлено в компактной форме:

$$\delta(\mathbf{r}) = 1 - \prod_{i=1}^{N} (1 - \delta_i(\mathbf{r})) \quad (85)$$

Объем молекулы может быть найден путем интегрирования этой индикаторной функции по пространству:

$$V = \iiint \delta(\mathbf{r}) d\mathbf{r} \quad (86)$$

Объем пересечения молекул A и B может в этом случае быть вычислен как интеграл перекрывания соответствующих им индикаторных функций $\delta_A(\mathbf{r})$ и $\delta_B(\mathbf{r})$:

$$V_{A,B} = \iiint \delta_A(\mathbf{r})\delta_B(\mathbf{r}) d\mathbf{r} \quad (87)$$

Этих величин вполне достаточно для вычисления рассмотренных выше количественных характеристик сходства пространственных форм молекул. Значения определенных интегралов в выражениях (86) и (87) можно оценить путем подсчета числа единичных значений индикаторной функции $\delta(\mathbf{r})$ на узлах равномерной и плотной решетки, охватывающей определенную область пространства. На этих принципах основана работа модуля CatShape [268], работающего в составе комплекса CATALYST и предназначенного для осуществления поиска по сходству пространственных форм молекул при проведении виртуального скрининга. В этом случае центр масс

таких точек образует центр молекулы, через который можно построить для них три главные оси инерции. Тогда объем молекулярной формы, определяемой этими точками, и длины отрезков, отсекаемых у осей инерции молекулярной поверхностью, определяемой этой формой, составляют четыре дескриптора молекулярной формы, которые могут быть использованы для отбрасывания на максимально ранней стадии молекул с формой, очень отличающейся от формы молекулы-шаблона. Далее центры масс молекул, наряду с главными осями, могут быть совмещены друг с другом, после чего положение и ориентация одной из молекул могут быть оптимизированы, начиная с этого положения, с целью максимизации перекрывания между молекулярными формами. После этого может быть по формуле (78) вычисляется значение индекса Танимото, который используется для количественной оценки сходства между формами двух молекул.

5.5.2.4. Сравнение пространственных форм молекул с использованием функций Гаусса

Альтернативный и существенно более эффективный в вычислительном плане по сравнению с моделью жестких сфер подход основан на описании форм молекул при помощи функций Гаусса [269, 270]. Огромное преимущество использования функций Гаусса для этой цели объясняется тем, что определенные интегралы как самих функций, так и их произведений, легко могут быть вычислены в аналитическом виде.

Идея описания пространственных форм молекул с использованием функций Гаусса заключается в замене индикаторных функций принадлежности точек атомам $\delta_i(\mathbf{r})$ на их «нечеткие» эквиваленты $g_i(\mathbf{r})$, построенные на основе сферических функций Гаусса:

$$g_i(\mathbf{r}) = p_i e^{-\alpha_i|\mathbf{r}-\mathbf{R}_i|^2} \qquad (88)$$

где p_i – параметр, равный значению функции Гаусса в максимуме в центре, т.е. на ядре атома i, α_i – параметр функции Гаусса, характеризующий скорость ее уменьшения при отдалении от центра. В этом случае «гауссовый» объем атома может быть найден путем интегрирования этой функции:

$$V = \iiint g_i(\mathbf{r})d\mathbf{r} = p_i \iiint e^{-\alpha_i|\mathbf{r}-\mathbf{R}_i|^2} d\mathbf{r} = p_i \left(\frac{\pi}{\alpha_i}\right)^{3/2} \qquad (89)$$

Если приравнять полученный по формуле (89) «гауссовый» объем атома вычисляемому по формуле (79) объему атома в модели жестких сфер, то легко прийти к условию равенства этих объемов:

$$\alpha_i = \pi \left(\frac{3p_i}{4\pi R_i^3} \right)^{2/3} \qquad (90)$$

Таким образом, если условие (90) удовлетворено, то объемы всех атомов, вычисляемые как в рамках модели жестких сфер, так и с помощью функций Гаусса, в точности совпадают. Остается, однако, свободный параметр p_i. Его значение принимают, исходя из требования, чтобы и объемы молекул, вычисляемые в рамках модели жестких сфер и с помощью функций Гаусса, совпадали. Объем молекулы по модели жестких сфер может быть оценен, например, как показано выше, с помощью метода Монте-Карло. Для вычисления же молекулярного объема при помощи функций Гаусса индикаторную функцию принадлежности точки молекуле $\delta(\mathbf{r})$ следует заменить на ее «нечеткий» аналог $\rho(\mathbf{r})$:

$$\rho(\mathbf{r}) = 1 - \prod_{i=1}^{N} (1 - g_i(\mathbf{r})) \qquad (91)$$

Интегрируя эту функцию по пространству, можно найти «гауссовый» объем молекулы. Хотя условия, обеспечивающего точное равенство объемов молекул, вычисляемых в рамках метода жестких сфер и с помощью функций Гаусса, по-видимому, не существует, тем не менее, в работе [269] эмпирически было найдено, что универсальное значение p=2.70 для всех атомов обеспечивает приблизительное равенство этих объемов с приемлемой точностью. Именно это значение параметра обычно и используется.

Огромное преимущество использования представления пространственных форм молекул при помощи функций Гаусса по сравнению с моделью жестких сфер заключается в возможности находить значения всех необходимых интегралов и их частных производных относительно координат атомов в аналитическом виде, что позволяет на порядки ускорять вычисления. Необходимые для этого вычислительные формулы приведены с статье [269]. Это дает возможность быстро вычислять интегралы перекрывания $O_{A,B}$ между функциями $\rho(\mathbf{r})$ для двух молекул:

$$O_{A,B} = \iiint \rho_A(\mathbf{r}) \rho_B(\mathbf{r}) d\mathbf{r} \qquad (92)$$

Используя этот интеграл перекрывания, можно вычислить значения мер сходства пространственных форм молекул, таких как

расстояние между формами ($D_{A,B}$), индекс Танимото ($T_{A,B}$) и индекс Тверского ($Tversky_{A,B}$) для пространственных форм молекул:

$$D_{A,B} = \sqrt{O_{A,A} + O_{B,B} - 2O_{A,B}} \quad (93)$$

$$T_{A,B} = \frac{O_{A,B}}{O_{A,A} + O_{B,B} - O_{A,B}} \quad (94)$$

$$Tversky_{A,B} = \frac{O_{A,B}}{\alpha O_{A,A} + \beta O_{B,B} - O_{A,B}} \quad (95)$$

Рассмотренная методология сравнения пространственных форм молекул с помощью функций Гаусса лежит в основе очень популярной программы для проведения виртуального скрининга ROCS (*Rapid Overlay of Chemical Structures*) [271], которая является продуктом фирмы OpenEye (https://www.eyesopen.com/rocs). В простейшем сценарии ее работы в качестве запроса при проведении виртуального скрининга задается пространственная структура активной молекулы, а поиск осуществляется в базе данных, в которых для каждой молекулы хранится представительный набор конформаций. Программа сравнивает пространственные формы молекул и осуществляет наложение молекулы запроса на молекулы из базы данных. При проведении сравнения и пространственного совмещения молекул функции Гаусса рассматриваются отдельно для разных химических (фармакофорных) типов атомов: доноров водородной связи, акцепторов водородной связи, анионов, катионов, гидрофобных атомов, циклов. В качестве количественных мер сходства используются индексы Танимото и Тверского. В последних версиях программы в качестве запроса может также фигурировать представленное с помощью решетки описание сайта связывания биологической мишени. Модифицированный вариант программы, FastROCS обеспечивает многократное ускорение работы благодаря использованию графических карт GPU. Программа ROCS (как и ее вариант FastROCS) обеспечивает очень быстрое и эффективное проведение виртуального скрининга, основанного на сравнении пространственных форм молекул.

5.5.2.5. Сравнение пространственных форм молекул с использованием сферических функций

Сферические функции, называемые также сферическими гармониками, часто используют для описания выраженных в сферических координатах угловых частей решений некоторых уравнений математической физики (в частности, уравнения Лапласа).

Их принято выражать в виде принимающей комплексные значения функции $Y_l^m(\theta,\varphi)$:

$$Y_l^m(\theta,\varphi) = \sqrt{\frac{2l+1}{4\pi}\cdot\frac{(l-m)!}{(l+m)!}}P_l^m(cos\theta)e^{-im\varphi} \qquad (96)$$

где: θ и ϕ – зенитный и азимутальный углы в сферической системе координат (r, θ, φ), см. Рис. 94; l и m – параметры функции; $P_l^m(cos\theta)$ – присоединенный многочлен Лежандра:

$$P_l^m(cos\theta) = sin^2\theta\frac{d^m}{d(cos\theta)^m}P_l(cos\theta) \qquad (97)$$

где $P_l(cos\theta)$ – многочлен Лежандра:

$$P_l(cos\theta) = \frac{1}{2^l l!}\cdot\frac{d^l}{d(cos\theta)^l}(cos^2\theta-1)^l \qquad (98)$$

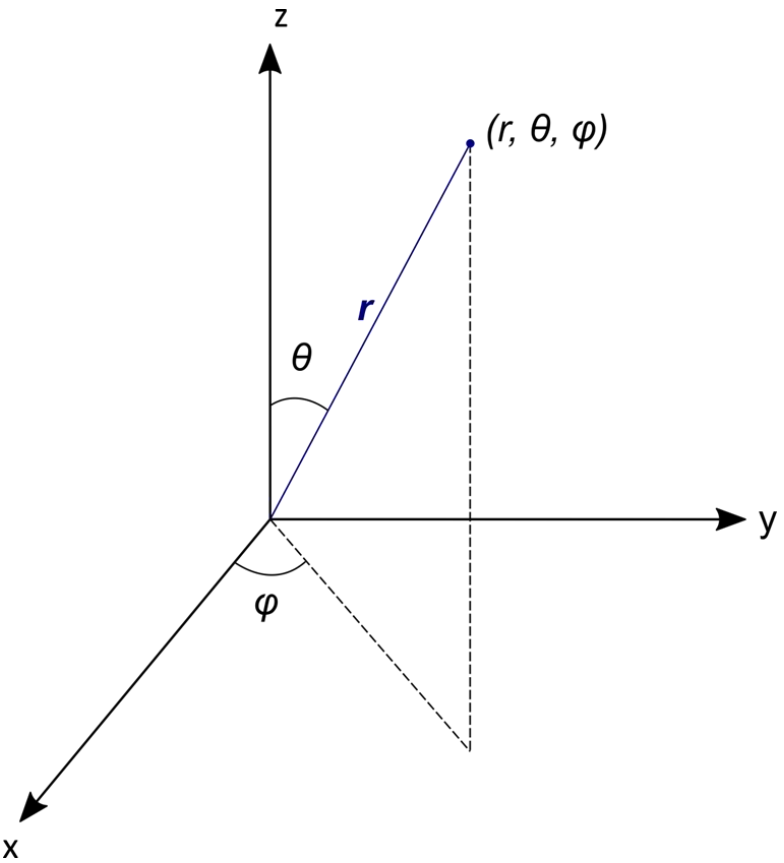


Рис. 94. Сферическая система координат (r, θ, φ): r – расстояние до начала координат; θ – зенитный угол; φ – азимутальный угол

Поскольку все геометрические параметры в физическом мире выражаются в действительных числах, вместо принимающих комплексные значения функций $Y_l^m(\theta,\varphi)$ удобно пользоваться их линейными комбинациями $S_l^m(\theta,\varphi)$, принимающими действительные значения:

$$S_l^m(\theta,\varphi) = \frac{1}{\sqrt{2}}\left(Y_l^m(\theta,\varphi) + Y_l^{-m}(\theta,\varphi)\right)$$
$$S_l^0(\theta,\varphi) = Y_l^0(\theta,\varphi) \qquad (99)$$
$$S_l^{-m}(\theta,\varphi) = \frac{1}{\sqrt{2}}\left(Y_l^m(\theta,\varphi) - Y_l^{-m}(\theta,\varphi)\right)$$

Самый известный пример использования функций $S_l^m(\theta,\varphi)$ – описание угловой части волновых функций атомов (т.е. атомных орбиталей). В этом случае параметр l является азимутальным квантовым числом, а параметр m – магнитным квантовым числом.

Важное свойство сферических функций состоит в том, что они образуют ортонормированный базис, по которому может быть разложена любая функция, использующая в качестве аргументов сферические координаты. Это означает, что если можно выбрать центр координат таким образом, чтобы расстояние r между ним и любой точкой на поверхности молекулы была однозначной функцией от сферических координат θ и φ, то эта функция может быть представлена как линейная комбинация базисных сферических функций:

$$r(\theta,\varphi) = \sum_{l=0}^{L}\sum_{m=-l}^{l} C_{lm} S_l^m(\theta,\varphi), \qquad (100)$$

где C_{lm} – набор коэффициентов разложения, который и определяет форму поверхности, а L – параметр, задающий степень сложности поверхности. При минимальном значении L=1 поверхность аппроксимируется в виде сферы, однако по мере возрастания L степень детализации поверхности возрастает, достигая максимально точной аппроксимации при больших значениях L.

Благодаря ортонормированности базиса сферических функций, значения коэффициентов разложения C_{lm} могут быть восстановлены из функции $r(\theta,\varphi)$, показывающей зависимость расстояния от точки на поверхности до начала координат в зависимости от углов θ и φ в сферической системе координат:

$$C_{lm} = \int_0^{\pi}\int_0^{2\pi} r(\theta,\varphi) S_l^m(\theta,\varphi) sin\theta d\theta d\varphi \qquad (101)$$

Заметим, что этим выражением можно воспользоваться только в том случае, когда функции $r(\theta,\varphi)$ является однозначной, т.е. одной комбинации значений углов θ и φ соответствует строго одно значение r. Это означает, что прямая линия, начинающаяся в начале координат и образующая с осями координат углы θ и φ (как показано на Рис. 94), пересекает поверхность строго в одной точке, как это показано на Рис. 95 на примере прямой 1. Такую поверхность можно считать «однозначной». В противном случае, при наличии нескольких точек

пересечения (прямая 2 на Рис. 95) такую поверхность можно считать «неоднозначной». Преобразования (101) и (100) позволяют однозначно переводить поверхность в набор коэффициентов и восстанавливать ее обратно только в том случае, если вся поверхность является «однозначной». Это можно считать ограничением рассматриваемого подхода. К счастью, для большинства молекул малого размера и даже для молекул большого размера при не очень большом значении L всегда можно задать такое начало координат, чтобы вся поверхность молекулы была «однозначной». В практическом плане вычисление двойного интеграла в формуле (101) может быть организовано по методу Монте-Карло путем проведении линий с равномерно распределенными случайными значениями углов θ и φ из начала координат, определения их точек пересечения с поверхностью и замены интегрирования на вычисление среднего значения.

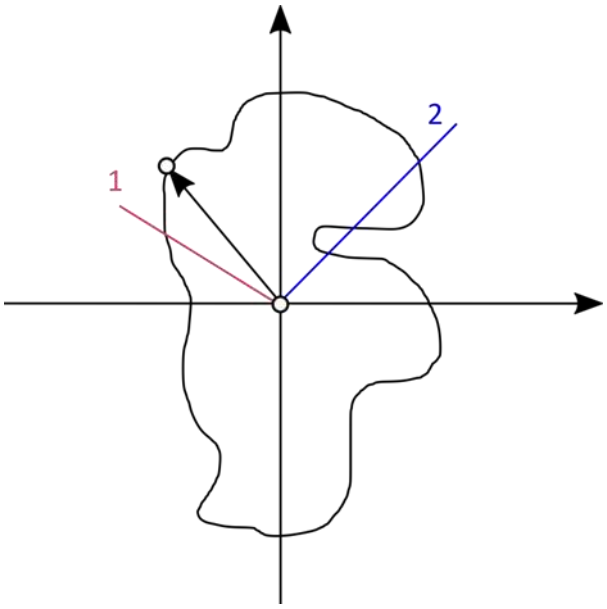


Рис. 95. Пояснение к понятию «однозначной» и «неоднозначной» поверхности. Линия, исходящая из начала координат, пересекает «однозначную» поверхность строго в одной точке (линия 1). Если же есть несколько точек пересечения (линия 2), то такая поверхность является «неоднозначной».

В основе рассматриваемого подхода к сравнению пространственных форм молекул лежит предположение о том, что это можно сделать путем сравнения *«однозначных» молекулярных поверхностей*, охватывающих эти формы. Это можно, например, сделать путем сравнения функций $r(\theta, \varphi)$, описывающих молекулы A и B с использованием общей системы сферических координат. В

частности, расстояние $D_{A,B}$ между ними может быть определено следующим образом [272]:

$$D_{A,B} = \sqrt{\int_0^{\pi}\int_0^{2\pi} \left(r_A(\theta,\varphi) - r_B(\theta,\varphi)\right)^2 sin\theta d\theta d\varphi}$$

$$= \sqrt{\int_0^{\pi}\int_0^{2\pi} \left(r_A(\theta,\varphi)^2 + r_B(\theta,\varphi)^2 - 2r_A(\theta,\varphi)r_B(\theta,\varphi)\right) sin\theta d\theta d\varphi} \quad (102)$$

$$= \sqrt{\sum_{l=0}^{L}\sum_{m=-l}^{l}(C_{lm}^A)^2 + \sum_{l=0}^{L}\sum_{m=-l}^{l}(C_{lm}^B)^2 - 2\sum_{l=0}^{L}\sum_{m=-l}^{l} C_{lm}^A C_{lm}^B} =$$

$$= \sqrt{\langle \mathbf{C}^A|\mathbf{C}^A\rangle + \langle \mathbf{C}^B|\mathbf{C}^B\rangle - 2\langle \mathbf{C}^A|\mathbf{C}^B\rangle}$$

где $\mathbf{C}^A$ обозначает вектор значений коэффициентов C_{lm}^A, собранный для допустимых при выбранном L значений индексов l и m, а $\langle \mathbf{C}^A|\mathbf{C}^B\rangle$ обозначает скалярное произведение векторов $\mathbf{C}^A$ и $\mathbf{C}^B$. Таким образом, расстояние между поверхностями, определяющими пространственную форму молекул, может быть выражено через скалярные произведения коэффициентов разложения поверхностей по сферическим функциям. Точно так же может быть определен и аналог индекса Танимото:

$$T_{A,B} = \frac{\langle \mathbf{C}^A|\mathbf{C}^B\rangle}{\langle \mathbf{C}^A|\mathbf{C}^A\rangle + \langle \mathbf{C}^B|\mathbf{C}^B\rangle - \langle \mathbf{C}^A|\mathbf{C}^B\rangle} \quad (103)$$

Формулы (102) и (103) могут быть использованы для организации виртуального скрининга путем сравнения форм молекул с помощью сферических функций. Вычисление по этим формулам требует выравнивания молекул A и B в пространстве при совмещенной точке начала координат. За нее может быть, например, принят центр масс молекулы. В этом случае совмещение молекул сводится к вращению одной из них. Это может быть достигнуто двумя способами. Первый из них сводится к минимизации расстояния $D_{A,B}$ между молекулами путем итерационного изменения угловой ориентации одной из молекул. Второй позволяет за один шаг ориентировать молекулы друг относительно друга путем совмещения их осей инерции. Второй способ является существенно более грубым, однако значительно более эффективным в вычислительном плане, что немаловажно при проведении виртуального скрининга.

При проведении виртуального скрининга за шаблон принимается поверхность активной молекулы, и осуществляется поиск в базе

данных молекул, поверхности которых наиболее к ней близки. При сравнении поверхностей можно учитывать распределение на них локальных свойств, например, электростатического потенциала. В этом случае при вычислении коэффициентов разложения вместо расстояния от центра координат используется значение локального свойства. В качестве вариантов возможно также использование в качестве шаблона «консенсусной» поверхности, сформированной путем усреднения поверхностей нескольких совмещенных в пространстве активных молекул [273]:

$$r^{cons}(\theta,\varphi) = \frac{1}{N_{cons}} \sum_{k=1}^{N_{cons}} \sum_{l=0}^{L} \sum_{m=-l}^{l} C_{lm} S_l^m(\theta,\varphi) \qquad (104)$$

где N_{cons} – число молекул, участвующих в консенсусном усреднении.

Теоретические основы изложенного подхода приведены в публикациях [272, 274, 275]. Примеры применения в виртуальном скрининге рассмотрены в публикациях [273, 276-278].

Описанные выше подходы реализованы в программах PARSURF (для построения и анализа молекулярных поверхностей) и PARAFIT (для выравнивания молекул в пространстве), являющимися коммерческими продуктами CEPOS InSilico Ltd., Erlangen, Germany, http://www.ceposinsilico.de.

5.5.2.6. Сравнение пространственных форм молекул с использованием статистических моментов наборов расстояний

Практически все рассмотренные выше подходы к сравнению пространственных форм молекул включают процедуру выравнивания молекул в пространстве, что часто приводит к большим затратам вычислительных ресурсов. Поэтому их нельзя применять для проведения виртуального скрининга реально существующих баз данных большого размера, включающих сотни миллионов конформаций молекул. Для решения этой проблемы были предложены подходы, основанные на сравнении статистических моментов распределений межатомных расстояний и других простейших геометрических характеристик, для вычисления которых не требуется проводить выравнивание молекул в пространстве. Хотя статистические параметры, характеризующие качество ранжирования, в таких подходах может быть несколько ниже по сравнению с методами, требующими выравнивания, однако чрезвычайно высокая

вычислительная эффективность делает их применение очень полезным по крайней мере на ранних этапах виртуального скрининга.

В 2007 г. был предложен метод «*сверхбыстрого распознавания формы*» (англ. Ultrafast Shape Recognition, *USR*), в котором пространственная форма молекул описывается при помощи дескрипторов – трех разных статистических моментов, описывающих распределение расстояний от всех атомов в молекуле до четырех специально выбранных ключевых точек (положений в пространстве) [279]. В качестве последних предложено использовать:

1. Центр молекулы (ctd);
2. Положение ближайшего атома к ctd (cst);
3. Положение атома, наиболее удаленного от ctd (fct);
4. Положение атома, наиболее удаленного от fct (ftf).

Это дает возможность для молекулы из N атомов рассчитать по N значений расстояний от каждого из атомов до каждого из этих четырех точек. Каждый из наборов из N расстояний d_i до каждой из этих четырех точек было предложено охарактеризовать первыми тремя статистическими моментами распределений:

1. Среднее (англ. mean): $\mu_1 = \frac{1}{N}\sum_{i=1}^{N} d_i$
2. Дисперсия (англ. variance): $\mu_2 = \frac{1}{N}\sum_{i=1}^{N}(d_i - \mu_1)^2$
3. Коэффициент асимметрии (англ. skewness): $\mu_3 = \frac{1}{N}\sum_{i=1}^{N}(d_i - \mu_1)^3$

Все это в совокупности дает набор из 12 дескрипторов: $\mathbf{m} = (\mu_1^{ctd}, \mu_2^{ctd}, \mu_3^{ctd}, \mu_1^{cst}, \mu_2^{cst}, \mu_3^{cst}, \mu_1^{fct}, \mu_2^{fct}, \mu_3^{fct}, \mu_1^{ftf}, \mu_2^{ftf}, \mu_3^{ftf})$. Таким образом, пространственная форма молекулы характеризуется вектором, состоящим из 12 дескрипторов. Следовательно, чтобы сравнить пространственные формы двух молекул, достаточно сравнить вычисленные для них таким образом вектора дескрипторов. В качестве меры сходства этих векторов было предложено использовать следующее выражение, значение которого лежит в интервале от 0 (очень сильно различающиеся пространственные формы молекул) до 1 (идентичные формы);

$$S_{A,B} = \frac{1}{1 + \frac{1}{12}\sum_{i=1}^{12}\left|m_i^A - m_i^B\right|} \qquad (105)$$

Эта величина, очевидно, представляет собой модификацию расстояния по Манхэттену (англ. Manhattan distance).

Несмотря на чрезвычайную простоту вычисления параметров, характеризующих пространственную форму молекулы, проводимый с их использованием виртуальный скрининг позволяет быстро находить

в базах данных очень большого размера молекулы с пространственной формой, очень близкой к желаемой (например, к пространственной форме молекулы, обладающей определенным видом биологической активности). Визуально можно отметить, что пространственная формы найденных таким образом молекул действительно очень близка к пространственной форме молекулы запроса.

У использования наборов межатомных расстояний для описания пространственных форм молекул имеется, однако, существенный недостаток – такие наборы не способны различать энантиомеры, поскольку при зеркальном отражении расстояние между точками не меняется. Этот недостаток был устранен путем перехода к альтернативному набору из четырех ключевых положений в рамках метода распознавания хиральной формы (англ. Chiral Form Recognition, *CSR*) [280]:

1. Геометрический центр молекулы ($\mathbf{p}_1$);
2. Положение атома, наиболее удаленного от $\mathbf{p}_1$ ($\mathbf{p}_2$);
3. Положение атома, наиболее удаленного от $\mathbf{p}_2$ ($\mathbf{p}_3$);
4. Вычисляется положение $\mathbf{p}_4$ по следующей формуле:

$$\mathbf{p}_4 = \mathbf{p}_1 + \frac{\|\mathbf{p}_2 - \mathbf{p}_1\|}{2} \cdot \frac{(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)}{\|(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)\|} \qquad (106)$$

где знак × обозначает векторное произведение, которое меняет знак при зеркальном отражении. В результате этого ключевое положение $\mathbf{p}_4$ будет разным для разных энантиомеров.

Дальнейшим развитием этого подходя явилось введение учета фармакофорных свойств атомов путем вычисления различных дескрипторов для атомов, относящихся к разным фармакофорным типам. В частности, в работе [281] представлен метод USRCAT, включающий вычисление различных моментов распределения для 5 разных фармакофорных типов:

1. Все атомы (o), дескрипторы с 1 по 12;
2. Гидрофобные атомы (h), дескрипторы с 13 по 24;
3. Атомы в составе ароматических систем (r), дескрипторы с 25 по 36;
4. Акцепторы (a), дескрипторы с 37 по 48;
5. Доноры (d), дескрипторы с 49 по 60.

В этом случае пространственная форма уже описывается при помощи 60 дескрипторов (по 12 на каждый фармакофорный тип), а при вычислении меры сходства пространственных форм моменты распределения для разных фармакофорных типов вводятся с разным весом:

$$S_{A,B}^{-1} = 1 + \frac{w^o}{12}\sum_{i=1}^{12}\left|m_i^A - m_i^B\right| + \frac{w^h}{12}\sum_{i=13}^{24}\left|m_i^A - m_i^B\right| + \frac{w^r}{12}\sum_{i=25}^{36}\left|m_i^A - m_i^B\right| + \frac{w^a}{12}\sum_{i=37}^{48}\left|m_i^A - m_i^B\right| + \frac{w^d}{12}\sum_{i=49}^{60}\left|m_i^A - m_i^B\right| \quad (107)$$

Рассмотренные выше методы виртуального скрининга, основанные на сравнении пространственных форм молекул при помощи статистических моментов наборов расстояний, реализованы на доступном через Интернет Web-сервере (http://usr.marseille.inserm.fr/) [282], позволяющем осуществить виртуальный скрининг встроенной базы данных большого размера, содержащей 93 миллиона низкоэнергетических конформаций, сгенерированных для 23 миллионов молекул, взятых из набора All Clean базы данных коммерчески доступных химических соединений ZINC.

5.5.3. Сходство молекулярных полей

В хемоинформатике используется более широкая по сравнению с физикой трактовка понятия поля. Под *молекулярным полем* понимается любая непрерывная функция f от пространственных координат **r**, описывающая межмолекулярные взаимодействия. Наиболее часто используются следующие типы молекулярных полей:

- Электростатические потенциал;
- Ван-дер-Ваальсовый (стерический) потенциал;
- Гидрофобный (липофильный) потенциал;
- Донорный потенциал водородной связи;
- Акцепторный потенциал водородной связи.

Легко заметить, что в представленном списке только электростатический потенциал является «настоящим» полем, рассматриваемым в физике, тогда как остальные представляют собой

сложные комбинации множества факторов, влияющих на взаимодействие между молекулами.

Использование молекулярных полей в виртуальном скрининге основано на предположении о том, что *лиганды со сходными молекулярными полями имеют тенденцию связываться с одними и теми же биологическими мишенями и, следовательно, проявлять сходную биологическую активность*. Следовательно, можно ожидать, что новые молекулы лекарств могут быть найдены с помощью виртуального скрининга на основе сходства молекулярных полей.

5.5.3.1. Оценка сходства полей с помощью интегралов перекрывания

Наиболее популярный метод оценки сходства молекулярных полей заключается в расчете интеграла их перекрывания (т.е. интеграла их произведения), как при оценке квантового сходства при помощи индекса Карбо (см. раздел 5.5.1). Принципиальное же отличие от оценки квантового сходства заключается в том, что корректные функциональные формы для электростатического и ван-дер-ваальсова потенциалов не могут быть проинтегрированы по всему пространству. С этой целью используют аппроксимацию этих полей при помощи одной или нескольких функций Гаусса, интеграл произведений которых всегда существует и, более того, может быть выражен в удобном для быстрых вычислений аналитическом виде.

Метод SEAL. Исторически первым подходом, основанным на этих принципах, явился метод «стерического и электронного выравнивания» SEAL (*Steric and Electronic ALignment*) [283], в котором для оценки сходства используется взвешенная линейная комбинация интегралов перекрывания электростатического $f^{el}(\mathbf{r})$ и стерического $f^{vdw}(\mathbf{r})$ полей молекул:

$$S_{A,B} = w^{el} \iiint f_A^{el}(\mathbf{r}) f_B^{el}(\mathbf{r}) d\mathbf{r} + w^{vdw} \iiint f_A^{vdw}(\mathbf{r}) f_B^{vdw}(\mathbf{r}) d\mathbf{r} \quad (108)$$

В рамках рассматриваемого подхода каждое из этих полей аппроксимируется при помощи суммы многомерных изотропных функций Гаусса, центрированных на ядрах атомов:

$$f^F(\mathbf{r}) = \sum_{i=1}^{N} exp(-\alpha(\mathbf{r} - \mathbf{R}_i)^2) \quad (109)$$

где $\mathbf{R}_i$ – радиус-вектор положения ядра i-ого атома в пространстве, N – число атомов в молекуле. Можно показать, что в этом случае индекс сходства полей двух молекул может быть представлен в следующем виде:

$$S_{A,B} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} w_{ij} exp\left(-\alpha d_{ij}^2\right) \quad (110)$$

где N_A – число атомов в молекуле A, N_B – число атомов в молекуле B, d_{ij} – геометрическое расстояние между атомом i в молекуле A и атомом j в молекуле B:

$$d_{ij}^2 = \left(\mathbf{R}_i^A - \mathbf{R}_j^B\right)^2 \quad (111)$$

Значение весового коэффициента w_{ij} задается путем комбинирования частичного заряда, q_i, и ван-дер-ваальсового радиуса атома i, R_i, в молекуле A и атома j в молекуле B:

$$w_{ij} = w^{el} q_i^A q_j^B + w^{vdw} R_i^A R_j^B \quad (112)$$

Как следует из формулы (110), значение индекса сходства $S_{A,B}$ зависит от набора расстояний между атомами молекул A и B. Следовательно, оно зависит от взаимной ориентации этих двух молекул. Поэтому для его вычисления проводят оптимизацию положения и ориентации одной молекулы относительно другой таким образом, чтобы значение индекса сходства стало максимальным:

$$S_{A,B} \rightarrow max \quad (113)$$

Это приводит к выравниванию молекул A и B в пространстве. Выравнивание молекул в пространстве является трудоемкой в вычислительном плане процедурой, что делает вычисление индекса сходства также трудоемким. Для ускорения этой процедуры часто используют методы оптимизации, основанные на алгебре кватернионов[1].

Метод FBSS. Дальнейшим развитием метода SEAL является метод FBSS (*Field-Based Similarity Searching*) [284-286]. Последний основан на использовании трех типов молекулярных полей: электростатического, стерического и гидрофобного, которые аппроксимируются при помощи многомерных изотропных функций Гаусса. Это позволяет получить более полное и точное описание полей, определяющих межмолекулярное взаимодействие, по сравнению с SEAL. Для выравнивания молекул в пространстве в методе FBSS используется генетический алгоритм, кодирующий в хромосомах

[1] См. подробнее https://ru.wikipedia.org/wiki/Кватернион

относительное положение и угловую ориентацию молекул из базы данных относительно молекулы запроса. Функция приспособленности (англ. fitness function) генетического алгоритма определяется мерой сходства молекулярных полей, вычисляемой с помощью интеграла перекрывания (индекса Карбо). Использование генетического алгоритма позволяет проводить пространственное совмещение в двух режимах: «жестком» (англ. rigid) и «гибком» (англ. flexible). В «жестком» режиме при выравнивании рассматривается только одна фиксированная конформация молекулы из базы данных, тогда как в «гибком» режиме одновременно ведется варьирование конформаций молекулы из базы данных благодаря включению значений ее гибких двугранных углов в хромосомы.

5.5.3.2. Оценка сходства полей с помощью полевых графов

Для ускорения процедуры совмещения молекул в пространстве было предложено описывать молекулярные поля при помощи *полевых графов* (англ. field graphs) [287], вершины которых соответствуют центрам кластеров узлов решетки со значением электростатического потенциала, превышающими пороговое, а ребра – всем их парам. В этом случае вершины могут быть помечены, например, средним значением электростатического потенциала на входящих в кластер узлах, а ребра – геометрическими расстояниями между узлами кластера. Тогда задача совмещения молекул в пространстве может быть сформулирована как задача поиска наибольшего общего подграфа (см. раздел 2.3.3 в пособии 2) для графов, характеризующих поля двух молекул. Решение этой задачи определяет соответствие между вершинами двух графов, что дает возможность их быстро наложить друг на друга в пространстве путем минимизации суммы квадратов расстояний между ними. Такое наложение позволяет найти в базе данных молекулы, электростатические поля которых похожи друг на друга.

В настоящее время наибольшей популярностью пользуются полевые графы, вершины которых соответствуют точкам с экстремальными значениями молекулярных полей различного типа (электростатического, стерического и гидрофобного) [288]. Подобные графы составляют основу методологии, реализованной в продуктах фирмы Cresset Biomolecular Discovery Ltd.[1] В этом случае поля любой молекулы описываются точками четырех типов, соответствующими:

[1] https://www.cresset-group.com/

(1) наиболее отрицательным значениям электростатического потенциала; (2) наиболее положительным значениям электростатического потенциала; (3) наиболее отрицательными значениями энергии ван-дер-Ваальса; (4) наибольшими значениями гидрофобного потенциала. Каждый из таких экстремумов изображается в виде шара, центр которого соответствует положению экстремума, а радиус – значению молекулярного поля в этом экстремуме.

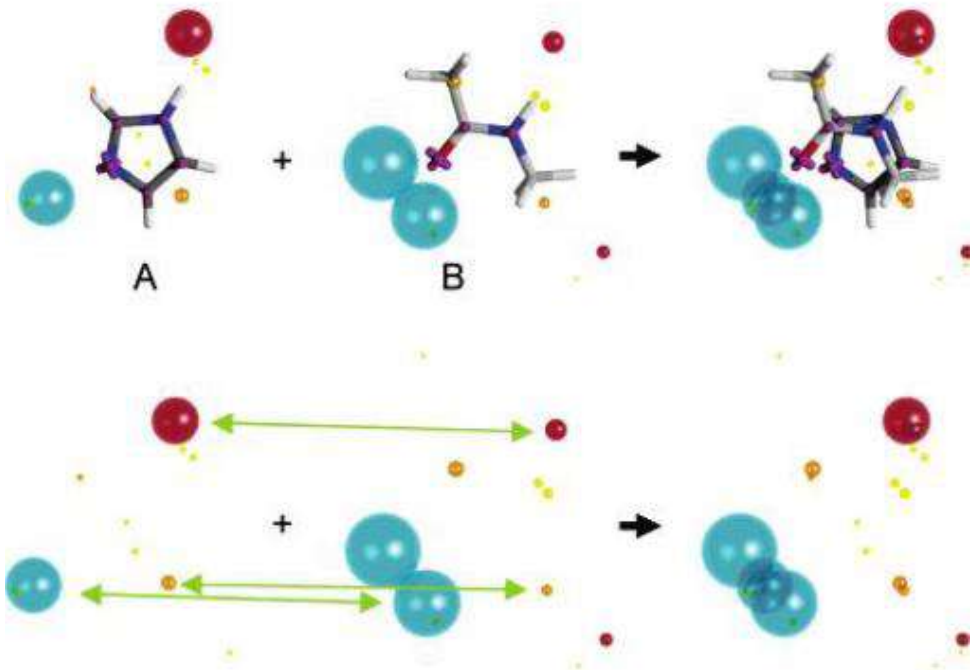


Рис. 96. Совмещение молекул в пространстве (сверху) путем совмещения точек, соответствующих экстремальным значениям полей (внизу). Точки голубого цвета соответствуют отрицательному электростатическому потенциалу, точки красного цвета – положительному, точки коричневого цвета – экстремумам гидрофобного потенциала, желтого цвета – экстремумам стерического потенциала. Стрелки обозначают соответствия между вершинами полевого графа, найденного с помощью процедуры поиска наиболее общее подграфа. Рисунок из статьи [288] публикуется с разрешения издательства. Copyright (2006) American Chemical Society.

На Рис. 96 приведены в качестве примера молекулы A и B вместе с экстремумами молекулярных полей. Наборы таких экстремумов образуют вершины полевых графов, тогда как ребра проведены между всеми парами вершин. Каждая вершина помечена типом экстремума, а ребро – геометрическим расстоянием между экстремумами.

Наибольший общий подграф для них содержит три вершины, которые соответствуют экстремумам для (1) отрицательного, (2) положительного значений электростатического потенциала, а также (3) экстремуму гидрофобного потенциала. Соответствие между этими тремя вершинами для обоих графов показано стрелками внизу. Такое соответствие позволяет очень быстро совместить экстремумы двух молекул друг с другом (внизу справа), что приводит к совмещению и самих молекул (вверху справа).

В качестве количественной меры сходства в рамках данного подхода было предложено использовать индекс Dice, выражающийся следующим образом:

$$S_{A,B} = \frac{E_{A\to B} + E_{B\to A}}{E_{A\to A} + E_{B\to B}} = \frac{2E_{A,B}}{E_{A,A} + E_{B,B}} \qquad (114)$$

где величина $E_{A\to B}$ имеет смысл энергии молекулы A в полях, создаваемых молекулой B:

$$E_{A\to B} = \sum_{fp_A} \sqrt{size(fp_A) \cdot F_B(position(fp_A))} \qquad (115)$$

где fp_A – точки экстремума (полевые точки), описывающие поля молекулы A, F_B – поле, создаваемое молекулой B.

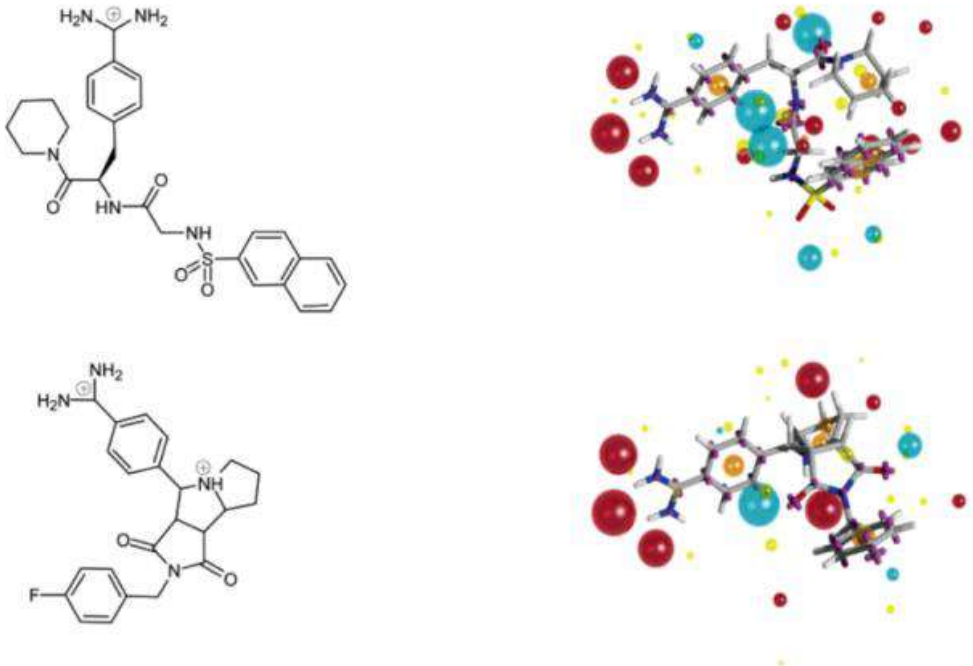


Рис. 97. Два ингибитора тромбина, принадлежащие разным классам химических соединений, но обладающие сходными молекулярными полями. Рисунок из статьи [288] публикуется с разрешения издательства. Copyright (2006) American Chemical Society.

В случае высоких значений индекса $S_{A,B}$ (т.е. приближающихся к единице) можно ожидать, что соответствующие молекулы будут связываться с одними и теми же биомишенями и, следовательно, обладать сходной биологической активностью, даже если чисто структурно они сильно отличаются друг от друга. На Рис. 96 в качестве примера приведены две молекулы, которые являются лигандами одного и того же фермента (тромбина). Хотя они принадлежат разным химическим классам, однако молекулярные поля, представленные с помощью экстремумов, у них очень похожи.

ЛИТЕРАТУРА

1. *A. Varnek, D. Fourches, F. Hoonakker, V.P. Solov'ev*. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. / A. Varnek, D. Fourches, F. Hoonakker, V.P. Solov'ev. // J. Comput. Aided Mol. Des. – 2005. – Т. 19, № 9-10. – С. 693-703.
2. Hypergraphs. North-Holland mathematical library. / C. Berge – Amsterdam: Elsevier, 1989. – Т. 45: North-Holland mathematical library.
3. Handbook of Molecular Descriptors. / R. Todeschini, V. Consonni – Weinheim: Wiley-VCH Publishers, 2000.
4. *N.M. Halberstam, I.I. Baskin, V.A. Palyulin, N.S. Zefirov*. Construction of neural-network structure-conditions-property relationships: Modeling of the physicochemical properties of hydrocarbons / N.M. Halberstam, I.I. Baskin, V.A. Palyulin, N.S. Zefirov. // Doklady Chemistry. – 2002. – Т. 384, № 1-3. – С. 140-143.
5. *D.J. Livingstone, D.W. Salt*. Variable selection - Spoilt for choice? / D.J. Livingstone, D.W. Salt. // Reviews in Computational Chemistry. – 2005. – Т. 21. – С. 287-348.
6. Pharmacophores and Pharmacophore Searches. / T. Langer, R.D. Hoffman – Weinheim: Wiley-VCH Publishers, 2000.
7. *J.M. Barnard*. A comparison of different approaches to Markush structure handling / J.M. Barnard. // Journal of Chemical Information and Computer Sciences. – 1991. – Т. 31, № 1. – С. 64-68.
8. *U. Schoch-Grübler*. (Sub)structure searches in databases containing generic chemical structure representations / U. Schoch-Grübler. // Online Information Review. – 1990. – Т. 14, № 2. – С. 95-108.
9. Concepts and Applications of Molecular Similarity. / A.M. Johnson, G.M. Maggiora – New York: John Willey & Sons, 1990.
10. *N. Nikolova, J. Jaworska*. Approaches to Measure Chemical Similarity - a Review / N. Nikolova, J. Jaworska. // QSAR & Combinatorial Science. – 2003. – Т. 22, № 9-10. – С. 1006-1026.
11. *J. Friedman, J.L. Bentley, R.A. Finkel*. An Algorithm for Finding Best Matches in Logarithmic Expected Time / J. Friedman, J.L. Bentley, R.A. Finkel. // ACM Transactions on Mathematics Software. – 1977. – Т. 3, № 3. – С. 209-226.
12. *P. Iyer, Y. Hu, J. Bajorath*. SAR Monitoring of Evolving Compound Data Sets Using Activity Landscapes / P. Iyer, Y. Hu, J. Bajorath. // Journal of Chemical Information and Modeling. – 2011. – Т. 51, № 3. – С. 532-540.
13. *J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse*. Reoptimization of MDL Keys for Use in Drug Discovery / J.L. Durant, B.A. Leland, D.R.

Henry, J.G. Nourse. // J. Chem. Inf. Comput. Sci. – 2002. – Т. 42, № 6. – С. 1273-1280.
14. *T.R. Hagadone*. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases / T.R. Hagadone. // J. Chem. Inf. Model. – 1992. – Т. 32, № 5. – С. 515-521.
15. *I.L. Ruiz, C.G. Garcia, M.A. Gomez-Nieto*. Clustering Chemical Databases Using Adaptable Projection Cells and MCS Similarity Values / I.L. Ruiz, C.G. Garcia, M.A. Gomez-Nieto. // J. Chem. Inf. Model. – 2005. – Т. 45, № 5. – С. 1178-1194.
16. *J.W. Raymond, P. Willett*. Maximum common subgraph isomorphism algorithms for the matching of chemical structures / J.W. Raymond, P. Willett. // J Comput Aided Mol Des. – 2002. – Т. 16, № 7. – С. 521-33.
17. *C. Bron, J. Kerbosch*. Algorithm 457: finding all cliques of an undirected graph / C. Bron, J. Kerbosch. // Commun. ACM. – 1973. – Т. 16. – С. 575-577.
18. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. / B. Schölkopf, A.J. Smola – Cambridge, MA; London, England: MIT Press, 2002.
19. *M. Rupp, R. Körner, I.V. Tetko*. Estimation of acid dissociation constants using graph kernels / M. Rupp, R. Körner, I.V. Tetko. // Molecular Informatics. – 2010. – Т. 29, № 10. – С. 731-740.
20. *M. Rupp, G. Schneider*. Graph Kernels for Molecular Similarity / M. Rupp, G. Schneider. // Molecular Informatics. – 2010. – Т. 29, № 4. – С. 266-273.
21. *P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert*. Graph Kernels for Molecular Structure−Activity Relationship Analysis with Support Vector Machines / P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert. // Journal of Chemical Information and Modeling. – 2005. – Т. 45, № 4. – С. 939-951.
22. *E. Bernard, Y. Jiao, E. Scornet, V. Stoven, T. Walter, J.-P. Vert*. Kernel Multitask Regression for Toxicogenetics / E. Bernard, Y. Jiao, E. Scornet, V. Stoven, T. Walter, J.-P. Vert. // Molecular Informatics. – 2017. – Т. 36, № 10. – С. 1700053-n/a.
23. *P. Jaccard*. Distribution de la flore alpine dans le Bassin des Dranses et dans quelque regions voisins / P. Jaccard. // Bull. Soc. Vaud. Sci. Nat. . – 1901. – Т. 37. – С. 241-272.
24. *S. Kulczinsky*. Zespoly róslin w Pienach / S. Kulczinsky. // Bull. intern. acad. polon. sci. lett. Cl. sci. math. natur. Ser. B. – 1927. – Т. Suppl. 2. – С. 57-203.
25. *D. Szymkiewicz*. Une contribution statistique a la géographie floristique / D. Szymkiewicz. // Acta Soc. Bot. Polon. – 1934. – Т. 34, № 3. – С. 249-265.

26. *G.G. Simpson*. Holarctic mammalian faunas and continental relationship during the Cenozoic / G.G. Simpson. // Bull. Geol. Sci. America. – 1947. – T. 58. – C. 613-688.
27. *J.D. Petke*. Cumulative and discrete similarity analysis of electrostatic potentials and fields / J.D. Petke. // Journal of Computational Chemistry. – 1993. – T. 14, № 8. – C. 928-933.
28. *A. Tversky*. Features of similarity / A. Tversky. // Physiological Reviews. – 1977. – T. 84, № 4. – C. 327-352.
29. *E.E. Hodgkin, W.G. Richards*. Molecular similarity based on electrostatic potential and electric field / E.E. Hodgkin, W.G. Richards. // International Journal of Quantum Chemistry. – 1987. – T. 32, № S14. – C. 105-110.
30. *B. Schölkopf, A. Smola, K.-R. Müller*. Nonlinear Component Analysis as a Kernel Eigenvalue Problem / B. Schölkopf, A. Smola, K.-R. Müller. // Neural Computation. – 1998. – T. 10, № 5. – C. 1299-1319.
31. *J.W. Sammon*. A Nonlinear Mapping for Data Structure Analysis / J.W. Sammon. // IEEE Trans. on Computer. – 1969. – T. 18. – C. 401-409.
32. *G.W. Bemis, M.A. Murcko*. The properties of known drugs. 1. Molecular frameworks / G.W. Bemis, M.A. Murcko. // J. Med. Chem. – 1996. – T. 39, № 15. – C. 2887-93.
33. *G.W. Bemis, M.A. Murcko*. Properties of known drugs. 2. Side chains / G.W. Bemis, M.A. Murcko. // J Med Chem. – 1999. – T. 42, № 25. – C. 5095-9.
34. *A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M.A. Koch, H. Waldmann*. The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification / A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M.A. Koch, H. Waldmann. // J Chem Inf Model. – 2007. – T. 47, № 1. – C. 47-58.
35. *M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann*. Charting biologically relevant chemical space: a structural classification of natural products (SCONP) / M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann. // Proc Natl Acad Sci U S A. – 2005. – T. 102, № 48. – C. 17272-7.
36. *S. Renner, W.A.L. Van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T.I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann*. Bioactivity-guided mapping and navigation of chemical space / S. Renner, W.A.L. Van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T.I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann. // Nature Chemical Biology. – 2009. – T. 5, № 8. – C. 585-592.

37. *S. Wetzel, K. Klein, S. Renner, D. Rauh, T.I. Oprea, P. Mutzel, H. Waldmann*. Interactive exploration of chemical space with Scaffold Hunter / S. Wetzel, K. Klein, S. Renner, D. Rauh, T.I. Oprea, P. Mutzel, H. Waldmann. // Nature Chemical Biology. – 2009. – Т. 5, № 8. – С. 581-583.
38. *L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath*. Exploration of structure-activity relationship determinants in analogue series / L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath. // Journal of Medicinal Chemistry. – 2009. – Т. 52, № 10. – С. 3212-3224.
39. *L. Peltason, J. Bajorath*. SAR Index: Quantifying the Nature of Structure-Activity Relationships / L. Peltason, J. Bajorath. // Journal of Medicinal Chemistry. – 2007. – Т. 50, № 23. – С. 5571-5578.
40. *D.K. Agrafiotis, M. Shemanarev, P.J. Connolly, M. Farnum, V.S. Lobanov*. SAR maps: A new SAR visualization technique for medicinal chemists / D.K. Agrafiotis, M. Shemanarev, P.J. Connolly, M. Farnum, V.S. Lobanov. // Journal of Medicinal Chemistry. – 2007. – Т. 50, № 24. – С. 5926-5937.
41. *S.N. Pollock, E.A. Coutsias, M.J. Wester, T.I. Oprea*. Scaffold topologies. 1. Exhaustive enumeration up to eight rings / S.N. Pollock, E.A. Coutsias, M.J. Wester, T.I. Oprea. // J. Chem. Inf. Mod. – 2008. – Т. 48, № 7. – С. 1304-1310.
42. *M.J. Wester, S.N. Pollock, E.A. Coutsias, T.K. Allu, S. Muresan, T.I. Oprea*. Scaffold topologies. 2. Analysis of chemical databases / M.J. Wester, S.N. Pollock, E.A. Coutsias, T.K. Allu, S. Muresan, T.I. Oprea. // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 7. – С. 1311-1324.
43. *E.V. Radchenko, V.A. Palyulin, N.S. Zefirov*. Molecular Field Topology Analysis in Drug Design and Virtual Screening // Chemoinformatics Approaches to Virtual Screening / Varnek A., Tropsha A.RSC, 2008. – С. 150-181.
44. *P.S. Magee*. A new Approach to Active-Site Binding Analysis. Inhibitors of Acetylcholinesterase / P.S. Magee. // Quantitative Structure-Activity Relationships. – 1990. – Т. 9, № 3. – С. 202-215.
45. *P.S. Magee*. Positional Analysis of Binding Events // QSAR: Rational Approaches to the Design of Bioactive Compounds / Silipo C., Vittoria A. – Amsterdam: Elsevier, 1991.
46. *C. Mercier, V. Fabart, Y. Sobel, J.E. Dubois*. Modeling alcohol metabolism with the DARC/CALPHI system / C. Mercier, V. Fabart, Y. Sobel, J.E. Dubois. // J Med Chem. – 1991. – Т. 34, № 3. – С. 934-42.
47. *L. Kurunczi, E. Seclaman, T.I. Oprea, L. Crisan, Z. Simon*. MTD-PLS: a PLS variant of the minimal topologic difference method. III. Mapping interactions between estradiol derivatives and the alpha estrogenic receptor

/ L. Kurunczi, E. Seclaman, T.I. Oprea, L. Crisan, Z. Simon. // J Chem Inf Model. – 2005. – T. 45, № 5. – C. 1275-81.
48. *L. Kurunczi, M. Olah, T.I. Oprea, C. Bologa, Z. Simon*. MTD-PLS: A PLS-based variant of the MTD method. 2. Mapping ligand-receptor interactions. Enzymatic acetic acid esters hydrolysis / L. Kurunczi, M. Olah, T.I. Oprea, C. Bologa, Z. Simon. // J Chem Inf Comput Sci. – 2002. – T. 42, № 4. – C. 841-6.
49. *T.I. Oprea, L. Kurunczi, M. Olah, Z. Simon*. MTD-PLS: a PLS-based variant of the MTD method. A 3D-QSAR analysis of receptor affinities for a series of halogenated dibenzoxin and biphenyl derivatives / T.I. Oprea, L. Kurunczi, M. Olah, Z. Simon. // SAR QSAR Environ Res. – 2001. – T. 12, № 1-2. – C. 75-92.
50. *V.A. Palyulin, E.V. Radchenko, N.S. Zefirov*. Molecular field topology analysis method in QSAR studies of organic compounds / V.A. Palyulin, E.V. Radchenko, N.S. Zefirov. // Journal of Chemical Information and Computer Sciences. – 2000. – T. 40, № 3. – C. 659-667.
51. *R. Van Deursen, J.L. Reymond*. Chemical space travel / R. Van Deursen, J.L. Reymond. // ChemMedChem. – 2007. – T. 2, № 5. – C. 636-640.
52. *K.J.M. Bishop, R. Klajn, B.A. Grzybowski*. The core and most useful molecules in organic chemistry / K.J.M. Bishop, R. Klajn, B.A. Grzybowski. // Angewandte Chemie - International Edition. – 2006. – T. 45, № 32. – C. 5348-5354.
53. *P.W. Kenny, J. Sadowski*. Structure Modification in Chemical Databases // Chemoinformatics in Drug DiscoveryWiley-VCH Verlag GmbH & Co. KGaA, 2005. – C. 271-285.
54. *E. Griffen, A.G. Leach, G.R. Robb, D.J. Warner*. Matched Molecular Pairs as a Medicinal Chemistry Tool / E. Griffen, A.G. Leach, G.R. Robb, D.J. Warner. // Journal of Medicinal Chemistry. – 2011. – T. 54, № 22. – C. 7739-7750.
55. *A.M. Wassermann, D. Dimova, P. Iyer, J. Bajorath*. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis / A.M. Wassermann, D. Dimova, P. Iyer, J. Bajorath. // Drug Development Research. – 2012. – T. 73, № 8. – C. 518-527.
56. *A.G. Dossetter, E.J. Griffen, A.G. Leach*. Matched Molecular Pair Analysis in drug discovery / A.G. Dossetter, E.J. Griffen, A.G. Leach. // Drug Discovery Today. – 2013. – T. 18, № 15-16. – C. 724-731.
57. *C. Tyrchan, E. Evertsson*. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations / C. Tyrchan, E. Evertsson. // Computational and Structural Biotechnology Journal. – 2017. – T. 15. – C. 86-90.

58. *A.G. Leach, H.D. Jones, D.A. Cosgrove, P.W. Kenny, L. Ruston, P. MacFaul, J.M. Wood, N. Colclough, B. Law*. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure / A.G. Leach, H.D. Jones, D.A. Cosgrove, P.W. Kenny, L. Ruston, P. MacFaul, J.M. Wood, N. Colclough, B. Law. // Journal of Medicinal Chemistry. – 2006. – Т. 49, № 23. – С. 6672-6682.
59. *Y. Sushko, S. Novotarskyi, R. Körner, J. Vogt, A. Abdelaziz, I.V. Tetko*. Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process / Y. Sushko, S. Novotarskyi, R. Körner, J. Vogt, A. Abdelaziz, I.V. Tetko. // Journal of Cheminformatics. – 2014. – Т. 6, № 1. – С. 48.
60. *J.W. Raymond, I.A. Watson, A. Mahoui*. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification / J.W. Raymond, I.A. Watson, A. Mahoui. // Journal of Chemical Information and Modeling. – 2009. – Т. 49, № 8. – С. 1952-1962.
61. *J.W. Raymond, E.J. Gardiner, P. Willett*. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm / J.W. Raymond, E.J. Gardiner, P. Willett. // Journal of Chemical Information and Computer Sciences. – 2002. – Т. 42, № 2. – С. 305-316.
62. *J.W. Raymond, P. Willett*. Maximum common subgraph isomorphism algorithms for the matching of chemical structures / J.W. Raymond, P. Willett. // Journal of Computer-Aided Molecular Design. – 2002. – Т. 16, № 7. – С. 521-533.
63. *J. Hussain, C. Rea*. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets / J. Hussain, C. Rea. // Journal of Chemical Information and Modeling. – 2010. – Т. 50, № 3. – С. 339-348.
64. *A.D. de Leon, J. Bajorath*. Matched molecular pairs derived by retrosynthetic fragmentation / A.D. de Leon, J. Bajorath. // Medchemcomm. – 2014. – Т. 5, № 1. – С. 64-67.
65. *X.Q. Lewell, D.B. Judd, S.P. Watson, M.M. Hann*. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry / X.Q. Lewell, D.B. Judd, S.P. Watson, M.M. Hann. // Journal of Chemical Information and Computer Sciences. – 1998. – Т. 38, № 3. – С. 511-522.
66. *N.M. O'Boyle, J. Boström, R.A. Sayle, A. Gill*. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity / N.M. O'Boyle,

J. Boström, R.A. Sayle, A. Gill. // Journal of Medicinal Chemistry. – 2014. – T. 57, № 6. – C. 2704-2713.
67. *E.S.R. Ehmki, C. Kramer*. Matched Molecular Series: Measuring SAR Similarity / E.S.R. Ehmki, C. Kramer. // Journal of Chemical Information and Modeling. – 2017. – T. 57, № 5. – C. 1187-1196.
68. *T.I. Oprea, J. Gottfries*. Chemography: The art of navigating in chemical space / T.I. Oprea, J. Gottfries. // Journal of Combinatorial Chemistry. – 2001. – T. 3, № 2. – C. 157-166.
69. *J. Larsson, J. Gottfries, L. Bohlin, A. Backlund*. Expanding the ChemGPS chemical space with natural products / J. Larsson, J. Gottfries, L. Bohlin, A. Backlund. // Journal of Natural Products. – 2005. – T. 68, № 7. – C. 985-991.
70. *J. Larsson, J. Gottfries, S. Muresan, A. Backlund*. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space / J. Larsson, J. Gottfries, S. Muresan, A. Backlund. // Journal of Natural Products. – 2007. – T. 70, № 5. – C. 789-794.
71. *A. Hyvarinen*. Independent component analysis: A neural network approach / A. Hyvarinen. // Acta Polytechnica Scandinavica Mathematics and Computing Series. – 1997. – T. 88.
72. *A. Hyvarinen, E. Oja*. A Fast Fixed-Point Algorithm for Independent Component Analysis / A. Hyvarinen, E. Oja. // Neural Computation. – 1997. – T. 9, № 7. – C. 1483-1492.
73. *A. Hyvarinen, E. Oja*. Independent component analysis: Algorithms and applications / A. Hyvarinen, E. Oja. // Neural Networks. – 2000. – T. 13, № 4-5. – C. 411-430.
74. *J. Chen, X.Z. Wang*. A new approach to near-infrared spectral data analysis using independent component analysis / J. Chen, X.Z. Wang. // Journal of Chemical Information and Computer Sciences. – 2001. – T. 41, № 4. – C. 992-1001.
75. *M.G. Gustafsson*. Independent component analysis yields chemically interpretable latent variables in multivariate regression / M.G. Gustafsson. // Journal of Chemical Information and Modeling. – 2005. – T. 45, № 5. – C. 1244-1255.
76. *D.K. Agrafiotis, D. Bandyopadhyay, M. Farnum*. Radial clustergrams: Visualizing the aggregate properties of hierarchical clusters / D.K. Agrafiotis, D. Bandyopadhyay, M. Farnum. // Journal of Chemical Information and Modeling. – 2007. – T. 47, № 1. – C. 69-75.
77. *A. Tropsha, D. Fourches*. Graph representation of molecular datasets: Applications to dataset visualization and comparison using graph indices / A. Tropsha, D. Fourches. // Chemistry Central Journal. – 2009. – T. 3, № SUPPL. 1.

78. *J. Hert, M.J. Keiser, J.J. Irwin, T.I. Oprea, B.K. Shoichet*. Quantifying the relationships among drug classes / J. Hert, M.J. Keiser, J.J. Irwin, T.I. Oprea, B.K. Shoichet. // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 4. – С. 755-765.
79. *M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath*. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices / M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath. // Journal of Medicinal Chemistry. – 2008. – Т. 51, № 19. – С. 6075-6084.
80. *L. Peltason, Y. Hu, J. Bajorath*. From structure-activity to structure-selectivity relationships: Quantitative assessment, selectivity cliffs, and key compounds / L. Peltason, Y. Hu, J. Bajorath. // ChemMedChem. – 2009. – Т. 4, № 11. – С. 1864-1873.
81. *Y.-H. Wang, Y. Li, S.-L. Yang, L. Yang*. Classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach / Y.-H. Wang, Y. Li, S.-L. Yang, L. Yang. // Journal of Chemical Information and Modeling. – 2005. – Т. 45, № 3. – С. 750-757.
82. *A. Ultsch, H.P. Siemon*. Kohonen's self-organizing feature maps for exploratory data analysis. // Book Kohonen's self-organizing feature maps for exploratory data analysis. / Editor, 1990. – С. 305-308.
83. *J. Iivarinen, T. Kohonen, J. Kangas, S. Kaski*. Visualizing the clusters on the self-organizing map. // Proceedings of the Conference on Artificial Intelligence Research in Finland / Carlsson C., Jaervi T., Reponen T. – Helsinki, Finland: Finnish Artificial Intelligence Society, 1994. – С. 122-126.
84. *M. von Korff, M. Steger*. GPCR-Tailored Pharmacophore Pattern Recognition of Small Molecular Ligands / M. von Korff, M. Steger. // Journal of Chemical Information and Computer Sciences. – 2004. – Т. 44, № 3. – С. 1137-1147.
85. *N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, A. Varnek*. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison / N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, A. Varnek. // Molecular Informatics. – 2012. – Т. 31, № 3-4. – С. 301-312.
86. *K. Klimenko, G. Marcou, D. Horvath, A. Varnek*. Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL Antiviral Compound Set / K. Klimenko, G. Marcou, D. Horvath, A. Varnek. // Journal of Chemical Information and Modeling. – 2016. – Т. 56, № 8. – С. 1438-1454.
87. *D.M. Maniyar, I.T. Nabney, B.S. Williams, A. Sewing*. Data Visualization during the Early Stages of Drug Discovery / D.M. Maniyar, I.T. Nabney,

B.S. Williams, A. Sewing. // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 4. – С. 1806-1818.
88. *P. Tino, I. Nabney*. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way / P. Tino, I. Nabney. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Т. 24, № 5. – С. 639-656.
89. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. / A. Lin, D. Horvath, V. Afonina, G. Marcou, J.-L. Reymond, A. Varnek, 2017.
90. *P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath*. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds / P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath. // Journal of Computer-Aided Molecular Design. – 2015. – Т. 29, № 12. – С. 1087-1108.
91. *R. Guha, J.H. Van Drie*. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs / R. Guha, J.H. Van Drie. // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 3. – С. 646-658.
92. *G.M. Maggiora*. On Outliers and Activity Cliffs Why QSAR Often Disappoints / G.M. Maggiora. // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 4. – С. 1535-1535.
93. *P.G. Polishchuk, T.I. Madzhidov, A. Varnek*. Estimation of the size of drug-like chemical space based on GDB-17 data / P.G. Polishchuk, T.I. Madzhidov, A. Varnek. // Journal of Computer-Aided Molecular Design. – 2013. – Т. 27, № 8. – С. 675-679.
94. *R.S. Bohacek, C. McMartin, W.C. Guida*. The art and practice of structure-based drug design: A molecular modeling perspective / R.S. Bohacek, C. McMartin, W.C. Guida. // Medicinal Research Reviews. – 1996. – Т. 16, № 1. – С. 3-50.
95. *W. Zheng, S.J. Cho, A. Tropsha*. Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries / W. Zheng, S.J. Cho, A. Tropsha. // Journal of Chemical Information and Computer Sciences. – 1998. – Т. 38, № 2. – С. 251-258.
96. *C.G. Wermuth*. Selective optimization of side activities: the SOSA approach / C.G. Wermuth. // Drug Discovery Today. – 2006. – Т. 11, № 3-4. – С. 160-164.
97. *C.-G. Wermuth*. Search for new lead compounds: The example of the chemical and pharmacological dissection of aminopyridazines / C.-G. Wermuth. // Journal of Heterocyclic Chemistry. – 1998. – Т. 35, № 5. – С. 1091-1100.

98. *C.-G. Wermuth*. Aminopyridazines - an alternative route to potent muscarinic agonists with no cholinergic syndrome / C.-G. Wermuth. // Il Farmaco. – 1993. – Т. 48. – С. 253-274.
99. *C.G. Wermuth, J.-J. Bourguignon, R.m. Hoffmann, R. Boigegrain, R. Brodin, J.-P. Kan, P. SoubriГ©*. SR 46559 A and related aminopyridazines are potent muscarinic agonists with no cholinergic syndrome / C.G. Wermuth, J.-J. Bourguignon, R.m. Hoffmann, R. Boigegrain, R. Brodin, J.-P. Kan, P. SoubriГ©. // Bioorganic & Medicinal Chemistry Letters. – 1992. – Т. 2, № 8. – С. 833-838.
100. *J. Degen, M. Rarey*. FlexNovo: Structure-Based Searching in Large Fragment Spaces / J. Degen, M. Rarey. // ChemMedChem. – 2006. – Т. 1, № 8. – С. 854-868.
101. FlexNovo // Book FlexNovo / EditorBioSolveIT GmbH, An der Ziegelei 75, 53757 St. Augustin, Germany, 2011.
102. *U. Fechner, G. Schneider*. Flux (2): Comparison of Molecular Mutation and Crossover Operators for Ligand-Based de Novo Design / U. Fechner, G. Schneider. // Journal of Chemical Information and Modeling. – 2007. – Т. 47, № 2. – С. 656-667.
103. *U. Fechner, G. Schneider*. Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design / U. Fechner, G. Schneider. // Journal of Chemical Information and Modeling. – 2005. – Т. 46, № 2. – С. 699-707.
104. *H.M. Vinkers, M.R. de Jonge, F.F.D. Daeyaert, J. Heeres, L.M.H. Koymans, J.H. van Lenthe, P.J. Lewi, H. Timmerman, K. Van Aken, P.A.J. Janssen*. SYNOPSIS: SYNthesize and OPtimize System in Silico / H.M. Vinkers, M.R. de Jonge, F.F.D. Daeyaert, J. Heeres, L.M.H. Koymans, J.H. van Lenthe, P.J. Lewi, H. Timmerman, K. Van Aken, P.A.J. Janssen. // Journal of Medicinal Chemistry. – 2003. – Т. 46, № 13. – С. 2765-2773.
105. *E.-W. Lameijer, J.N. Kok, T. BIГ¤ck, A.P. Ijzerman*. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules / E.-W. Lameijer, J.N. Kok, T. BIГ¤ck, A.P. Ijzerman. // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 2. – С. 545-552.
106. *D. Douguet, H.l. Munier-Lehmann, G. Labesse, S. Pochet*. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design / D. Douguet, H.l. Munier-Lehmann, G. Labesse, S. Pochet. // Journal of Medicinal Chemistry. – 2005. – Т. 48, № 7. – С. 2457-2468.
107. ReCore, part of LeadIT 2.1 // Book ReCore, part of LeadIT 2.1 / Editor. – St. Augustin: BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany, 2012. – С. Fast Core Replacement.

108. BROOD // Book BROOD / Editor. – Santa Fe, New Mexico, USA: OpenEye Scientific Software, 1997-2010. – C. Fragment Replacement and Molecular Design.
109. SparkV10 // Book SparkV10 / Editor. – Welwyn Garden: Cresset Group, BioPark Hertfordshire, Broadwater Road, Welwyn Garden City, Hertfordshire AL7 3AX, United Kingdom, 2012. – C. Exciting and powerful way of generating novel and diverse structures for your project.
110. *M. Congreve, R. Carr, C. Murray, H. Jhoti*. A Rule of Three for fragment-based lead discovery? / M. Congreve, R. Carr, C. Murray, H. Jhoti. // Drug Discovery Today. – 2003. – T. 8, № 19. – C. 876-877.
111. *X.Q. Lewell, D.B. Judd, S.P. Watson, M.M. Hann*. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry / X.Q. Lewell, D.B. Judd, S.P. Watson, M.M. Hann. // Journal of Chemical Information and Computer Sciences. – 1998. – T. 38, № 3. – C. 511-522.
112. *M. Waldman, H. Li, M. Hassan*. Novel algorithms for the optimization of molecular diversity of combinatorial libraries / M. Waldman, H. Li, M. Hassan. // Journal of Molecular Graphics and Modelling. – 2000. – T. 18, № 4вЂ“5. – C. 412-426.
113. *P. Willett*. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds / P. Willett. // Journal of Computational Biology. – 1999. – T. 6, № 3-4. – C. 447-457.
114. *V.J. Gillet, P. Willett*. Dissimilarity-based compound selection for library design // Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery / Ghose A. K., Viswanadhan V. N. – New York: Marcel Dekker, 2001. – C. 379-398.
115. *R.W. Kennard, L.A. Stone*. Computer aided design of experiments / R.W. Kennard, L.A. Stone. // Technometrics. – 1969. – C. 137-148.
116. *J.D. Holliday, P. Willett*. Definitions of "Dissimilarity" for Dissimilarity-Based Compound Selection / J.D. Holliday, P. Willett. // Journal of Biomolecular Screening. – 1996. – T. 1, № 3. – C. 145-151.
117. *M. Snarey, N.K. Terrett, P. Willett, D.J. Wilton*. Comparison of algorithms for dissimilarity-based compound selection / M. Snarey, N.K. Terrett, P. Willett, D.J. Wilton. // Journal of Molecular Graphics and Modelling. – 1997. – T. 15, № 6. – C. 372-385.
118. *R.D. Clark*. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets / R.D. Clark. // Journal of Chemical Information and Computer Sciences. – 1997. – T. 37, № 6. – C. 1181-1188.
119. *J.D. Holliday, S.S. Ranade, P. Willett*. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases / J.D.

Holliday, S.S. Ranade, P. Willett. // Quantitative Structure-Activity Relationships. – 1995. – T. 14, № 6. – C. 501-506.
120. *D.K. Agrafiotis, V.S. Lobanov*. An Efficient Implementation of Distance-Based Diversity Measures Based on k-d Trees / D.K. Agrafiotis, V.S. Lobanov. // Journal of Chemical Information and Computer Sciences. – 1998. – T. 39, № 1. – C. 51-58.
121. *B.D. Hudson, R.M. Hyde, E. Rahr, J. Wood, J. Osman*. Parameter Based Methods for Compound Selection from Chemical Databases / B.D. Hudson, R.M. Hyde, E. Rahr, J. Wood, J. Osman. // Quantitative Structure-Activity Relationships. – 1996. – T. 15, № 4. – C. 285-289.
122. *R. Wootton, R. Cranfield, G.C. Sheppey, P.J. Goodford*. Physicochemical-activity relations in practice. 2. Rational selection of benzenoid substituents / R. Wootton, R. Cranfield, G.C. Sheppey, P.J. Goodford. // Journal of Medicinal Chemistry. – 1975. – T. 18, № 6. – C. 607-613.
123. SYBYL-X Suite // Book SYBYL-X Suite / EditorTripos International: 1699 South Hanley Road, St. Louis, MO 63144-2319 USA, 2012. – C. Molecular Modeling from Sequence through Lead Optimization.
124. *R.A. Lewis, J.S. Mason, I.M. McLay*. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries:вЂ‰ The Diverse Property-Derived (DPD) Approach / R.A. Lewis, J.S. Mason, I.M. McLay. // Journal of Chemical Information and Computer Sciences. – 1997. – T. 37, № 3. – C. 599-614.
125. *D.J. Cummins, C.W. Andrews, J.A. Bentley, M. Cory*. Molecular Diversity in Chemical Databases:вЂ‰ Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available CompoundsвЂ / D.J. Cummins, C.W. Andrews, J.A. Bentley, M. Cory. // Journal of Chemical Information and Computer Sciences. – 1996. – T. 36, № 4. – C. 750-763.
126. *M.J. Bayley, P. Willett*. Binning schemes for partition-based compound selection / M.J. Bayley, P. Willett. // Journal of Molecular Graphics and Modelling. – 1999. – T. 17, № 1. – C. 10-18.
127. *D.K. Agrafiotis*. Stochastic Algorithms for Maximizing Molecular Diversity / D.K. Agrafiotis. // Journal of Chemical Information and Computer Sciences. – 1997. – T. 37, № 5. – C. 841-851.
128. *M. Hassan, J. Bielawski, J. Hempel, M. Waldman*. Optimization and visualization of molecular diversity of combinatorial libraries / M. Hassan, J. Bielawski, J. Hempel, M. Waldman. // Molecular Diversity. – 1996. – T. 2, № 1. – C. 64-74.
129. *D.K. Agrafiotis*. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries / D.K. Agrafiotis. // Journal of

Chemical Information and Computer Sciences. – 2001. – Т. 41, № 1. – С. 159-167.
130. *A.C. Good, R.A. Lewis*. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick / A.C. Good, R.A. Lewis. // Journal of Medicinal Chemistry. – 1997. – Т. 40, № 24. – С. 3926-3936.
131. *R.D. Brown, M. Hassan, M. Waldman*. Combinatorial library design for diversity, cost efficiency, and drug-like character / R.D. Brown, M. Hassan, M. Waldman. // Journal of Molecular Graphics and Modelling. – 2000. – Т. 18, № 4вЂ“5. – С. 427-437.
132. *D.K. Agrafiotis*. Multiobjective optimization of combinatorial libraries / D.K. Agrafiotis. // Journal of Computer-Aided Molecular Design. – 2002. – Т. 16, № 5-6. – С. 335-356.
133. *М.Д. Андреев, Д.Г. Хороших*. Многокритериальная оптимизация в аспекте антикризисного управления / М.Д. Андреев, Д.Г. Хороших. // Антикризисное управление. – 2002. – Т. 11-12.
134. *V.J. Gillet, P. Willett, P.J. Fleming, D.V.S. Green*. Designing focused libraries using MoSELECT / V.J. Gillet, P. Willett, P.J. Fleming, D.V.S. Green. // Journal of Molecular Graphics and Modelling. – 2002. – Т. 20, № 6. – С. 491-498.
135. *M. Lebl*. Parallel Personal Comments on Classical Papers in Combinatorial Chemistry / M. Lebl. // Journal of Combinatorial Chemistry. – 1998. – Т. 1, № 1. – С. 3-24.
136. Combinatorial Chemistry: From Theory to Application / Сост. Mannhold R., Kubinyi H., Folkers G. – Weinheim: Wiley-VCH, 2006. – 672 с.
137. Combinatorial Chemistry. / N.K. Terrett – New York: Oxford University Press, 1998. – 186 с.
138. *X.-D. Xiang, X. Sun, G. BriceГ±o, Y. Lou, K.-A. Wang, H. Chang, W.G. Wallace-Freedman, S.-W. Chen, P.G. Schultz*. A Combinatorial Approach to Materials Discovery / X.-D. Xiang, X. Sun, G. BriceГ±o, Y. Lou, K.-A. Wang, H. Chang, W.G. Wallace-Freedman, S.-W. Chen, P.G. Schultz. // Science. – 1995. – Т. 268, № 5218. – С. 1738-1740.
139. *E. Danielson, J.H. Golden, E.W. McFarland, C.M. Reaves, W.H. Weinberg, X.D. Wu*. A combinatorial approach to the discovery and optimization of luminescent materials / E. Danielson, J.H. Golden, E.W. McFarland, C.M. Reaves, W.H. Weinberg, X.D. Wu. // Nature. – 1997. – Т. 389, № 6654. – С. 944-948.
140. *A. Schüller, V. Hähnke, G. Schneider*. SmiLib v2.0: A Java-Based Tool for Rapid Combinatorial Library Enumeration / A. Schüller, V. Hähnke, G.

Schneider. // QSAR & Combinatorial Science. – 2007. – Т. 26, № 3. – С. 407-410.
141. *G.M. Downs, J.M. Barnard*. Techniques for Generating Descriptive Fingerprints in Combinatorial LibrariesB§ / G.M. Downs, J.M. Barnard. // Journal of Chemical Information and Computer Sciences. – 1997. – Т. 37, № 1. – С. 59-61.
142. Torus // Book Torus / EditorDigital Chemistry, 30 Kiveton Lane, Todwick, Sheffield S26 1HL, United Kingdom 2012.
143. *E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos*. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery / E.J. Martin, J.M. Blaney, M.A. Siani, D.C. Spellmeyer, A.K. Wong, W.H. Moos. // Journal of Medicinal Chemistry. – 1995. – Т. 38, № 9. – С. 1431-1436.
144. *V.J. Gillet, P. Willett, J. Bradshaw*. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries / V.J. Gillet, P. Willett, J. Bradshaw. // Journal of Chemical Information and Computer Sciences. – 1997. – Т. 37, № 4. – С. 731-740.
145. *E.A. Jamois, M. Hassan, M. Waldman*. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets / E.A. Jamois, M. Hassan, M. Waldman. // Journal of Chemical Information and Computer Sciences. – 2000. – Т. 40, № 1. – С. 63-70.
146. *W. Zheng, S.T. Hung, J.T. Saunders, G.L. Seibel*. PICCOLO: a tool for combinatorial library design via multicriterion optimization / W. Zheng, S.T. Hung, J.T. Saunders, G.L. Seibel. // Pac Symp Biocomput. – 2000. – С. 588-99.
147. *V.J. Gillet, P. Willett, J. Bradshaw, D.V.S. Green*. Selecting Combinatorial Libraries to Optimize Diversity and Physical Properties / V.J. Gillet, P. Willett, J. Bradshaw, D.V.S. Green. // Journal of Chemical Information and Computer Sciences. – 1999. – Т. 39, № 1. – С. 169-177.
148. *R.P. Sheridan, S.G. SanFeliciano, S.K. Kearsley*. Designing targeted libraries with genetic algorithms / R.P. Sheridan, S.G. SanFeliciano, S.K. Kearsley. // Journal of Molecular Graphics and Modelling. – 2000. – Т. 18, № 4ВЂ“5. – С. 320-334.
149. *R.P. Sheridan, S.K. Kearsley*. Using a Genetic Algorithm To Suggest Combinatorial Libraries / R.P. Sheridan, S.K. Kearsley. // Journal of Chemical Information and Computer Sciences. – 1995. – Т. 35, № 2. – С. 310-320.
150. *G. Liang, S. Aldous, G. Merriman, J. Levell, J. Pribish, J. Cairns, X. Chen, S. Maignan, M. Mathieu, J. Tsay, K. Sides, S. Rebello, B. Whitely, I. Morize, H.W. Pauls*. Structure-based library design and the discovery of a potent and selective mast cell beta-tryptase inhibitor as an oral therapeutic

agent / G. Liang, S. Aldous, G. Merriman, J. Levell, J. Pribish, J. Cairns, X. Chen, S. Maignan, M. Mathieu, J. Tsay, K. Sides, S. Rebello, B. Whitely, I. Morize, H.W. Pauls. // Bioorganic & Medicinal Chemistry Letters. – 2012. – Т. 22, № 2. – С. 1049-1054.
151. *E.K. Kick, D.C. Roe, A.G. Skillman, G. Liu, T.J.A. Ewing, Y. Sun, I.D. Kuntz, J.A. Ellman*. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D / E.K. Kick, D.C. Roe, A.G. Skillman, G. Liu, T.J.A. Ewing, Y. Sun, I.D. Kuntz, J.A. Ellman. // Chemistry and Biology. – 1997. – Т. 4, № 4. – С. 297-307.
152. *H.J. Bohm, M. Stahl*. Structure-based library design: Molecular modelling merges with combinatorial chemistry / H.J. Bohm, M. Stahl. // Current Opinion in Chemical Biology. – 2000. – Т. 4, № 3. – С. 283-286.
153. *D. Roe, I. Kuntz*. BUILDER v.2: Improving the chemistry of a de novo design strategy / D. Roe, I. Kuntz. // Journal of Computer-Aided Molecular Design. – 1995. – Т. 9, № 3. – С. 269-282.
154. *T.S. Haque, A.G. Skillman, C.E. Lee, H. Habashita, I.Y. Gluzman, T.J.A. Ewing, D.E. Goldberg, I.D. Kuntz, J.A. Ellman*. Potent, Low-Molecular-Weight Non-Peptide Inhibitors of Malarial Aspartyl Protease Plasmepsin II / T.S. Haque, A.G. Skillman, C.E. Lee, H. Habashita, I.Y. Gluzman, T.J.A. Ewing, D.E. Goldberg, I.D. Kuntz, J.A. Ellman. // Journal of Medicinal Chemistry. – 1999. – Т. 42, № 8. – С. 1428-1440.
155. *H.-J. Bohm, D.W. Banner, L. Weber*. Combinatorial docking and combinatorial chemistry: Design of potent non-peptide thrombin inhibitors / H.-J. Bohm, D.W. Banner, L. Weber. // Journal of Computer-Aided Molecular Design. – 1999. – Т. 13, № 1. – С. 51-56.
156. *M.G. Bursavich, C.W. West, D.H. Rich*. From Peptides to Non-Peptide Peptidomimetics:вЂ‰ Design and Synthesis of New Piperidine Inhibitors of Aspartic Peptidases / M.G. Bursavich, C.W. West, D.H. Rich. // Organic Letters. – 2001. – Т. 3, № 15. – С. 2317-2320.
157. *W.E. Minke, F. Hong, C.L.M.J. Verlinde, W.G.J. Hol, E. Fan*. Using a Galactose Library for Exploration of a Novel Hydrophobic Pocket in the Receptor Binding Site of the Escherichia coli Heat-labile Enterotoxin / W.E. Minke, F. Hong, C.L.M.J. Verlinde, W.G.J. Hol, E. Fan. // Journal of Biological Chemistry. – 1999. – Т. 274, № 47. – С. 33469-33473.
158. *A. Lew, A.R. Chamberlin*. Blockers of human T cell Kv1.3 potassium channels using de novo ligand design and solid-phase parallel combinatorial chemistry / A. Lew, A.R. Chamberlin. // Bioorganic & Medicinal Chemistry Letters. – 1999. – Т. 9, № 23. – С. 3267-3272.
159. *C.M. Huwe*. Synthetic library design / C.M. Huwe. // Drug Discovery Today. – 2006. – Т. 11, № 15вЂ“16. – С. 763-767.

160. *A.M. ter Laak, J. Venhorst, G.M. Donne-Op den Kelder, H. Timmerman*. The Histamine H1-Receptor Antagonist Binding Site. A Stereoselective Pharmacophoric Model Based upon (Semi-)Rigid H1-Antagonists and Including a Known Interaction Site on the Receptor / A.M. ter Laak, J. Venhorst, G.M. Donne-Op den Kelder, H. Timmerman. // Journal of Medicinal Chemistry. – 1995. – Т. 38, № 17. – С. 3351-3360.
161. *G. Wolber, T. Langer*. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters / G. Wolber, T. Langer. // Journal of Chemical Information and Modeling. – 2004. – Т. 45, № 1. – С. 160-169.
162. LigandScout 3.0 // Book LigandScout 3.0 / EditorInte:Ligand, Software-Entwicklungs und Consulting GmbH, Clemens Maria Hofbauer-G. 6, A-2344 Maria Enzersdorf Austria, Europe 2012. – С. Advanced Pharmacophore Modeling.
163. *F. Ortuso, T. Langer, S. Alcaro*. GBPM: GRID-based pharmacophore model: concept and application studies to proteinвЂ“protein recognition / F. Ortuso, T. Langer, S. Alcaro. // Bioinformatics. – 2006. – Т. 22, № 12. – С. 1449-1455.
164. *H.A. Carlson, K.M. Masukawa, K. Rubins, F.D. Bushman, W.L. Jorgensen, R.D. Lins, J.M. Briggs, J.A. McCammon*. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase / H.A. Carlson, K.M. Masukawa, K. Rubins, F.D. Bushman, W.L. Jorgensen, R.D. Lins, J.M. Briggs, J.A. McCammon. // Journal of Medicinal Chemistry. – 2000. – Т. 43, № 11. – С. 2100-2114.
165. *I. Motoc, R.A. Dammkoehler, D. Mayer, J. Labanowski*. Three-Dimensional Quantitative Structure-Activity Relationships I. General Approach to the Pharmacophore Model Validation / I. Motoc, R.A. Dammkoehler, D. Mayer, J. Labanowski. // Quantitative Structure-Activity Relationships. – 1986. – Т. 5, № 3. – С. 99-105.
166. *G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler, D.A. Dunn*. The conformational parameter in drug design: the active analogue approach in computer-assisted drug design // Computer-Assisted Drug Design / Olson E. C., Christoffersen R. E. – Washington DC: American Chemical Society, 1979. – С. 205-226.
167. *G. Jones, P. Willett, R.C. Glen*. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation / G. Jones, P. Willett, R.C. Glen. // Journal of Computer-Aided Molecular Design. – 1995. – Т. 9, № 6. – С. 532-549.
168. GASP // Book GASP / EditorTripos International: 1699 South Hanley Road, St. Louis, MO 63144-2319 USA, 2012. – С. Develop Pharmacophore Hypotheses Using Full Conformational Flexibility.

169. *A. Strizhev, E.J. Abrahamian, S. Choi, J.M. Leonard, P.R.N. Wolohan, R.D. Clark*. The Effects of Biasing Torsional Mutations in a Conformational GA / A. Strizhev, E.J. Abrahamian, S. Choi, J.M. Leonard, P.R.N. Wolohan, R.D. Clark. // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 4. – С. 1862-1870.
170. *G. Jones, P. Willett, R.C. Glen*. GASP: Genetic Algorythm Superimposition Program // Pharmacophore Perception, Development, and Use in Drug Design / Guner O. F. – La Jolla, California, USA: International University Line, 2000. – С. 85-107.
171. *N. Richmond, C. Abrams, P. Wolohan, E. Abrahamian, P. Willett, R. Clark*. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D / N. Richmond, C. Abrams, P. Wolohan, E. Abrahamian, P. Willett, R. Clark. // Journal of Computer-Aided Molecular Design. – 2006. – Т. 20, № 9. – С. 567-587.
172. GALAHAD // Book GALAHAD / EditorTripos International: 1699 South Hanley Road, St. Louis, MO 63144-2319 USA, 2012. – C. Rapid, High Quality Pharmacophoric Perception and Molecular Alignments.
173. *S.J. Cottrell, V.J. Gillet, R. Taylor, D.J. Wilton*. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques / S.J. Cottrell, V.J. Gillet, R. Taylor, D.J. Wilton. // Journal of Computer-Aided Molecular Design. – 2004. – Т. 18, № 11. – С. 665-682.
174. *S. Cottrell, V. Gillet, R. Taylor*. Incorporating partial matches within multiobjective pharmacophore identification / S. Cottrell, V. Gillet, R. Taylor. // Journal of Computer-Aided Molecular Design. – 2006. – Т. 20, № 12. – С. 735-749.
175. *E.J. Gardiner, D.A. Cosgrove, R. Taylor, V.J. Gillet*. Multiobjective Optimization of Pharmacophore Hypotheses: Bias Toward Low-Energy Conformations / E.J. Gardiner, D.A. Cosgrove, R. Taylor, V.J. Gillet. // Journal of Chemical Information and Modeling. – 2009. – Т. 49, № 12. – С. 2761-2773.
176. *G. Jones*. GAPE: An Improved Genetic Algorithm for Pharmacophore Elucidation / G. Jones. // Journal of Chemical Information and Modeling. – 2010. – Т. 50, № 11. – С. 2001-2018.
177. *C. Bron, J. Kerbosch*. Algorithm 457: finding all cliques of an undirected graph / C. Bron, J. Kerbosch. // Commun. ACM. – 1973. – Т. 16, № 9. – С. 575-577.
178. *Y.C. Martin, M.G. Bures, E.A. Danaher, J. DeLazzer, I. Lico, P.A. Pavlik*. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists / Y.C. Martin, M.G. Bures, E.A. Danaher, J. DeLazzer, I. Lico, P.A. Pavlik. // Journal of Computer-Aided Molecular Design. – 1993. – Т. 7, № 1. – С. 83-102.

179. *Y.C. Martin*. DISCO: What We Did Right and What We Missed // Pharmacophore Perception, Development, and Use in Drug Design / Guner O. F. – La Jolla, California, USA: International University Line, 2000. – С. 49-69.
180. DISCOtech // Book DISCOtech / EditorTripos International: 1699 South Hanley Road, St. Louis, MO 63144-2319 USA, 2012. – С. Rapid, High Quality Pharmacophoric Perception and Molecular Alignments.
181. *D. Barnum, J. Greene, A. Smellie, P. Sprague*. Identification of Common Functional Configurations Among Molecules / D. Barnum, J. Greene, A. Smellie, P. Sprague. // Journal of Chemical Information and Computer Sciences. – 1996. – Т. 36, № 3. – С. 563-571.
182. Discovery Studio Modeling Environment // Book Discovery Studio Modeling Environment / Editor. – San Diego: Accelrys Software Inc., 2012., 2012.
183. *A. Smellie, S.L. Teig, P. Towbin*. Poling: Promoting conformational variation / A. Smellie, S.L. Teig, P. Towbin. // Journal of Computational Chemistry. – 1995. – Т. 16, № 2. – С. 171-187.
184. *A. Smellie, S.D. Kahn, S.L. Teig*. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage / A. Smellie, S.D. Kahn, S.L. Teig. // Journal of Chemical Information and Computer Sciences. – 1995. – Т. 35, № 2. – С. 285-294.
185. *J. Greene, S. Kahn, H. Savoj, P. Sprague, S. Teig*. Chemical Function Queries for 3D Database Search / J. Greene, S. Kahn, H. Savoj, P. Sprague, S. Teig. // Journal of Chemical Information and Computer Sciences. – 1994. – Т. 34, № 6. – С. 1297-1308.
186. Phase // Book Phase / EditorSchrodinger, 2012.
187. *G. Wolber, A. Dornhofer, T. Langer*. Efficient overlay of small organic molecules using 3D pharmacophores / G. Wolber, A. Dornhofer, T. Langer. // Journal of Computer-Aided Molecular Design. – 2006. – Т. 20, № 12. – С. 773-788.
188. OMEGA // Book OMEGA / Editor. – Santa Fe, New Mexico, USA: OpenEye Scientific Software, 1997-2012. – С. Conformer Ensembles Containing Bioactive Conformations.
189. *H.W. Kuhn*. The Hungarian method for the assignment problem / H.W. Kuhn. // Naval Research Logistics Quarterly. – 1955. – Т. 2, № 1-2. – С. 83-97.
190. *H.W. Kuhn*. Variants of the hungarian method for assignment problems / H.W. Kuhn. // Naval Research Logistics Quarterly. – 1956. – Т. 3, № 4. – С. 253-258.

191. *W. Kabsch*. A solution for the best rotation to relate two sets of vectors / W. Kabsch. // Acta Crystallographica Section A. – 1976. – Т. 32, № 5. – С. 922-923.
192. *W. Kabsch*. A discussion of the solution for the best rotation to relate two sets of vectors / W. Kabsch. // Acta Crystallographica Section A. – 1978. – Т. 34, № 5. – С. 827-828.
193. *P.G. Polishchuk, G.V. Samoylenko, T.M. Khristova, O.L. Krysko, T.A. Kabanova, V.M. Kabanov, A.Y. Kornylov, O. Klimchuk, T. Langer, S.A. Andronati, V.E. Kuz'min, A.A. Krysko, A. Varnek*. Design, Virtual Screening, and Synthesis of Antagonists of αIIbβ3 as Antiplatelet Agents / P.G. Polishchuk, G.V. Samoylenko, T.M. Khristova, O.L. Krysko, T.A. Kabanova, V.M. Kabanov, A.Y. Kornylov, O. Klimchuk, T. Langer, S.A. Andronati, V.E. Kuz'min, A.A. Krysko, A. Varnek. // Journal of Medicinal Chemistry. – 2015. – Т. 58, № 19. – С. 7681-7694.
194. *J.F. Truchon, C.I. Bayly*. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem / J.F. Truchon, C.I. Bayly. // Journal of Chemical Information and Modeling. – 2007. – Т. 47, № 2. – С. 488-508.
195. *Q. Hu, Z. Peng, S.C. Sutton, J. Na, J. Kostrowicki, B. Yang, T. Thacher, X. Kong, S. Mattaparti, J.Z. Zhou, J. Gonzalez, M. Ramirez-Weinhouse, A. Kuki*. Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information / Q. Hu, Z. Peng, S.C. Sutton, J. Na, J. Kostrowicki, B. Yang, T. Thacher, X. Kong, S. Mattaparti, J.Z. Zhou, J. Gonzalez, M. Ramirez-Weinhouse, A. Kuki. // ACS Combinatorial Science. – 2012. – Т. 14, № 11. – С. 579-589.
196. *C. Lipinski*. Computational alerts for potential absorption problems: profiles of clinically tested drugs // Book Computational alerts for potential absorption problems: profiles of clinically tested drugs / Editor. – Miami, 1995.
197. *C.A. Lipinski*. Drug-like properties and the causes of poor solubility and poor permeability / C.A. Lipinski. // Journal of Pharmacological and Toxicological Methods. – 2000. – Т. 44, № 1. – С. 235-249.
198. *A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski*. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases / A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski. // Journal of Combinatorial Chemistry. – 1999. – Т. 1, № 1. – С. 55-68.
199. *D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple*. Molecular properties that influence the oral bioavailability of drug candidates / D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple. // J. Med. Chem. – 2002. – Т. 45, № 12. – С. 2615-23.

200. *M. Congreve, R. Carr, C. Murray, H. Jhoti*. A 'rule of three' for fragment-based lead discovery? / M. Congreve, R. Carr, C. Murray, H. Jhoti. // Drug Discovery Today. – 2003. – Т. 8, № 19. – С. 876-877.
201. *R. Benigni, C. Bossa, O. Tcheremenskaia*. Nongenotoxic Carcinogenicity of Chemicals: Mechanisms of Action and Early Recognition through a New Set of Structural Alerts / R. Benigni, C. Bossa, O. Tcheremenskaia. // Chemical Reviews. – 2013. – Т. 113, № 5. – С. 2940-2957.
202. *J.L. Hermens*. Electrophiles and acute toxicity to fish / J.L. Hermens. // Environmental Health Perspectives. – 1990. – Т. 87. – С. 219-225.
203. *H.J.M. Verhaar, C.J. van Leeuwen, J.L.M. Hermens*. Classifying environmental pollutants / H.J.M. Verhaar, C.J. van Leeuwen, J.L.M. Hermens. // Chemosphere. – 1992. – Т. 25, № 4. – С. 471-491.
204. *M.D. Barratt, D.A. Basketter, M. Chamberlain, G.D. Admans, J.J. Langowski*. An expert system rulebase for identifying contact allergens / M.D. Barratt, D.A. Basketter, M. Chamberlain, G.D. Admans, J.J. Langowski. // Toxicology in Vitro. – 1994. – Т. 8, № 5. – С. 1053-1060.
205. *I. Gerner, M.D. Barratt, S. Zinke, K. Schlegel, E. Schlede*. Development and prevalidation of a list of structure-activity relationship rules to be used in expert systems for prediction of the skin-sensitising properties of chemicals / I. Gerner, M.D. Barratt, S. Zinke, K. Schlegel, E. Schlede. // Alternatives to laboratory animals : ATLA. – 2004. – Т. 32, № 5. – С. 487-509.
206. *M.P. Payne, P.T. Walsh*. Structure-activity relationships for skin sensitization potential: Development of structural alerts for use in knowledge-based toxicity prediction systems / M.P. Payne, P.T. Walsh. // Journal of Chemical Information and Computer Sciences. – 1994. – Т. 34, № 1. – С. 154-161.
207. *S.J. Enoch, J.C. Madden, M.T.D. Cronin*. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach / S.J. Enoch, J.C. Madden, M.T.D. Cronin. // SAR and QSAR in Environmental Research. – 2008. – Т. 19, № 5-6. – С. 555-578.
208. *J. Kazius, R. McGuire, R. Bursi*. Derivation and validation of toxicophores for mutagenicity prediction / J. Kazius, R. McGuire, R. Bursi. // J. Med. Chem. – 2005. – Т. 48, № 1. – С. 312-20.
209. *R. Benigni, C. Bossa*. Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology / R. Benigni, C. Bossa. // Mutation Research/Reviews in Mutation Research. – 2008. – Т. 659, № 3. – С. 248-261.
210. *A.B. Bailey, R. Chanderbhan, N. Collazo-Braier, M.A. Cheeseman, M.L. Twaroski*. The use of structure–activity relationship analysis in the food

contact notification program / A.B. Bailey, R. Chanderbhan, N. Collazo-Braier, M.A. Cheeseman, M.L. Twaroski. // Regulatory Toxicology and Pharmacology. – 2005. – Т. 42, № 2. – С. 225-235.
211. *J. Ashby, R.W. Tennant*. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP / J. Ashby, R.W. Tennant. // Mutation Research/Genetic Toxicology. – 1988. – Т. 204, № 1. – С. 17-115.
212. *A. Plošnik, M. Vračko, M. Sollner Dolenc*. Mutagenic and carcinogenic structural alerts and their mechanisms of action / A. Plošnik, M. Vračko, M. Sollner Dolenc. // Arhiv za higijenu rada i toksikologiju. – 2016. – Т. 67, № 3. – С. 169-182.
213. *A.S. Kalgutkar, J.R. Soglia*. Minimising the potential for metabolic activation in drug discovery / A.S. Kalgutkar, J.R. Soglia. // Expert Opinion on Drug Metabolism & Toxicology. – 2005. – Т. 1, № 1. – С. 91-142.
214. *J.L. Dahlin, J.W.M. Nissink, J.M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M.A. Walters*. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS / J.L. Dahlin, J.W.M. Nissink, J.M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M.A. Walters. // Journal of Medicinal Chemistry. – 2015. – Т. 58, № 5. – С. 2091-2113.
215. *M.F. Sassano, A.K. Doak, B.L. Roth, B.K. Shoichet*. Colloidal Aggregation Causes Inhibition of G Protein-Coupled Receptors / M.F. Sassano, A.K. Doak, B.L. Roth, B.K. Shoichet. // Journal of Medicinal Chemistry. – 2013. – Т. 56, № 6. – С. 2406-2414.
216. *N. Thorne, D.S. Auld, J. Inglese*. Apparent activity in high-throughput screening: origins of compound-dependent assay interference / N. Thorne, D.S. Auld, J. Inglese. // Current Opinion in Chemical Biology. – 2010. – Т. 14, № 3. – С. 315-324.
217. *J.B. Baell, G.A. Holloway*. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays / J.B. Baell, G.A. Holloway. // Journal of Medicinal Chemistry. – 2010. – Т. 53, № 7. – С. 2719-2740.
218. *D. Lagorce, O. Sperandio, J.B. Baell, M.A. Miteva, B.O. Villoutreix*. FAF-Drugs3: a web server for compound property calculation and chemical library design / D. Lagorce, O. Sperandio, J.B. Baell, M.A. Miteva, B.O. Villoutreix. // Nucleic Acids Research. – 2015. – Т. 43, № W1. – С. W200-W207.
219. *S.J. Capuzzi, E.N. Muratov, A. Tropsha*. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS / S.J.

Capuzzi, E.N. Muratov, A. Tropsha. // Journal of Chemical Information and Modeling. – 2017. – T. 57, № 3. – C. 417-427.
220. *D. Lagorce, N. Oliveira, M.A. Miteva, B.O. Villoutreix*. Pan-assay interference compounds (PAINS) that may not be too painful for chemical biology projects / D. Lagorce, N. Oliveira, M.A. Miteva, B.O. Villoutreix. // Drug Discovery Today. – 2017. – T. 22, № 8. – C. 1131-1133.
221. *E. Ahlberg, L. Carlsson, S. Boyer*. Computational Derivation of Structural Alerts from Large Toxicology Data Sets / E. Ahlberg, L. Carlsson, S. Boyer. // Journal of Chemical Information and Modeling. – 2014. – T. 54, № 10. – C. 2945-2952.
222. *I. Cortes-Ciriano*. Bioalerts: a python library for the derivation of structural alerts from bioactivity and toxicity data sets / I. Cortes-Ciriano. // Journal of Cheminformatics. – 2016. – T. 8. – C. 13.
223. *D.M. Sanderson, C.G. Earnshaw*. Computer prediction of possible toxic action from chemical structure; the DEREK system / D.M. Sanderson, C.G. Earnshaw. // Hum. Exp. Toxicol. – 1991. – T. 10, № 4. – C. 261-73.
224. *P.N. Judson*. Rule Induction for Systems Predicting Biological Activity / P.N. Judson. // J. Chem. Inf. Comput. Sci. – 1994. – T. 34, № 1. – C. 148-153.
225. *N. Greene, P.N. Judson, J.J. Langowski, C.A. Marchant*. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR / N. Greene, P.N. Judson, J.J. Langowski, C.A. Marchant. // SAR and QSAR in Environmental Research. – 1999. – T. 10, № 2-3. – C. 299-+.
226. *R. Benigni, C. Bossa*. Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens / R. Benigni, C. Bossa. // Toxicol. Mech. Methods. – 2008. – T. 18, № 2-3. – C. 137-147.
227. *A. Lepailleur, G. Poezevara, R. Bureau*. Automated detection of structural alerts (chemical fragments) in (eco)toxicology / A. Lepailleur, G. Poezevara, R. Bureau. // Computational and Structural Biotechnology Journal. – 2013. – T. 5. – C. e201302013.
228. *M. Floris, G. Raitano, R. Medda, E. Benfenati*. Fragment Prioritization on a Large Mutagenicity Dataset / M. Floris, G. Raitano, R. Medda, E. Benfenati. // Molecular Informatics. – 2017. – T. 36, № 7. – C. 1600133-n/a.
229. *H. Yang, J. Li, Z. Wu, W. Li, G. Liu, Y. Tang*. Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark / H. Yang, J. Li, Z. Wu, W. Li, G. Liu, Y. Tang. // Chemical Research in Toxicology. – 2017. – T. 30, № 6. – C. 1355-1364.
230. *G. Klopman*. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of

organic molecules / G. Klopman. // J. Am. Chem. Soc. – 1984. – T. 106, № 24. – C. 7315-21.
231. *G. Klopman, H.S. Rosenkranz*. Structural requirements for the mutagenicity of environmental nitroarenes / G. Klopman, H.S. Rosenkranz. // Mutat. Res. – 1984. – T. 126, № 3. – C. 227-38.
232. *G. Klopman*. MULTICASE. 1. A Hierarchical computer automated structure evaluation program / G. Klopman. // Quant. Struct.-Act. Relat. – 1992. – T. 11, № 2. – C. 176-84.
233. *G. Klopman*. The MultiCASE Program II. Baseline Activity Identification Algorithm (BAIA) / G. Klopman. // Journal of Chemical Information and Computer Sciences. – 1998. – T. 38, № 1. – C. 78-81.
234. *A.R. Cunningham, S.T. Moss, S.A. Iype, G. Qian, S. Qamar, S.L. Cunningham*. Structure−Activity Relationship Analysis of Rat Mammary Carcinogens / A.R. Cunningham, S.T. Moss, S.A. Iype, G. Qian, S. Qamar, S.L. Cunningham. // Chemical Research in Toxicology. – 2008. – T. 21, № 10. – C. 1970-1982.
235. *T. Ferrari, D. Cattaneo, G. Gini, N. Golbamaki Bakhtyari, A. Manganaro, E. Benfenati*. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction / T. Ferrari, D. Cattaneo, G. Gini, N. Golbamaki Bakhtyari, A. Manganaro, E. Benfenati. // SAR and QSAR in Environmental Research. – 2013. – T. 24, № 5. – C. 365-383.
236. *A. Golbamaki, E. Benfenati, N. Golbamaki, A. Manganaro, E. Merdivan, A. Roncaglioni, G. Gini*. New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds / A. Golbamaki, E. Benfenati, N. Golbamaki, A. Manganaro, E. Merdivan, A. Roncaglioni, G. Gini. // Journal of Environmental Science and Health, Part C. – 2016. – T. 34, № 2. – C. 97-113.
237. *C.W. Yap*. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints / C.W. Yap. // Journal of Computational Chemistry. – 2011. – T. 32, № 7. – C. 1466-1474.
238. *J. Klekota, F.P. Roth*. Chemical substructures that enrich for biological activity / J. Klekota, F.P. Roth. // Bioinformatics. – 2008. – T. 24, № 21. – C. 2518-2525.
239. *C. Helma, T. Cramer, S. Kramer, L. De Raedt*. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds / C. Helma, T. Cramer, S. Kramer, L. De Raedt. // Journal of Chemical Information and Computer Sciences. – 2004. – T. 44, № 4. – C. 1402-1411.
240. *L. De Raedt, S. Kramer*. The Levelwise Version Space Algorithm and its Application to Molecular Fragment Finding // The Seventeenth

International Joint Conference on Articial Intelligence. – San Francisco: Morgan Kaufmann, 2001. – C. 853-862.
241. *Y. Chi, R.R. Muntz, S. Nijssen, J.N. Kok*. Frequent subtree mining -- an overview. / Y. Chi, R.R. Muntz, S. Nijssen, J.N. Kok. // Fundamenta Informaticae –2005. – T. 66, № 1-2. – C. 161-198.
242. *J. Han, H. Cheng, D. Xin, X. Yan*. Frequent pattern mining: current status and future directions / J. Han, H. Cheng, D. Xin, X. Yan. // Data Mining and Knowledge Discovery. – 2007. – T. 15, № 1. – C. 55-86.
243. *C. Borgelt, T. Meinl, M. Berthold*. MoSS: A Program for Molecular Substructure Mining // Proceedings of the 1st international Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations –New York, NY: ACM Press, 2005. – C. 6-15.
244. *S. Nijssen, J.N. Kok*. A quickstart in frequent structure mining can make a difference // Book A quickstart in frequent structure mining can make a difference / Editor. – Seattle, WA, USA: ACM, 2004. – C. 647-652.
245. *X.F. Yan, J.W. Han*. gSpan: Graph-based substructure pattern mining // 2002 IEEE International Conference on Data Mining, Proceedings / Kumar V., Tsumoto S., Zhong N., Yu P. S., Wu X. D. – Los Alamitos: IEEE Computer Soc., 2002. – C. 721-724.
246. *G. Dong, J. Li*. Efficient mining of emerging patterns: discovering trends and differences // Book Efficient mining of emerging patterns: discovering trends and differences / Editor. – San Diego, California, USA: ACM, 1999. – C. 43-52.
247. *J. Auer, J. Bajorath*. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection / J. Auer, J. Bajorath. // Journal of Chemical Information and Modeling. – 2006. – T. 46, № 6. – C. 2502-2514.
248. *S. Lozano, G. Poezevara, M.-P. Halm-Lemeille, E. Lescot-Fontaine, A. Lepailleur, R. Bissell-Siders, B. Crémilleux, S. Rault, B. Cuissart, R. Bureau*. Introduction of Jumping Fragments in Combination with QSARs for the Assessment of Classification in Ecotoxicology / S. Lozano, G. Poezevara, M.-P. Halm-Lemeille, E. Lescot-Fontaine, A. Lepailleur, R. Bissell-Siders, B. Crémilleux, S. Rault, B. Cuissart, R. Bureau. // Journal of Chemical Information and Modeling. – 2010. – T. 50, № 8. – C. 1330-1339.
249. *B. Cuissart, G. Poezevara, B. Crémilleux, A. Lepailleur, R. Bureau*. Emerging Patterns as Structural Alerts for Computational Toxicology // Contrast Data MiningJames Bailey, 2013. – C. chapitre 19.
250. *R. Sherhod, V.J. Gillet, P.N. Judson, J.D. Vessey*. Automating Knowledge Discovery for Toxicity Prediction Using Jumping Emerging Pattern Mining / R. Sherhod, V.J. Gillet, P.N. Judson, J.D. Vessey. // Journal

of Chemical Information and Modeling. – 2012. – T. 52, № 11. – C. 3074-3087.
251. *J.P. Metivier, A. Lepailleur, A. Buzmakov, G. Poezevara, B. Cremileux, S.O. Kuznetsov, J. Le Goff, A. Napoli, R. Bureau, B. Cuissart*. Discovering Structural Alerts for Mutagenicity Using Stable Emerging Molecular Patterns / J.P. Metivier, A. Lepailleur, A. Buzmakov, G. Poezevara, B. Cremileux, S.O. Kuznetsov, J. Le Goff, A. Napoli, R. Bureau, B. Cuissart. // Journal of Chemical Information and Modeling. – 2015. – T. 55, № 5. – C. 925-940.
252. *A.A. Khalifa, M. Haranczyk, J. Holliday*. Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection / A.A. Khalifa, M. Haranczyk, J. Holliday. // Journal of Chemical Information and Modeling. – 2009. – T. 49, № 5. – C. 1193-1201.
253. *P. Willett*. Combination of Similarity Rankings Using Data Fusion / P. Willett. // Journal of Chemical Information and Modeling. – 2013. – T. 53, № 1. – C. 1-10.
254. *J.W. Raymond, M. Jalaie, M.P. Bradley*. Conditional Probability: A New Fusion Method for Merging Disparate Virtual Screening Results / J.W. Raymond, M. Jalaie, M.P. Bradley. // Journal of Chemical Information and Computer Sciences. – 2004. – T. 44, № 2. – C. 601-609.
255. *J.C. Baber, W.A. Shirley, Y. Gao, M. Feher*. The Use of Consensus Scoring in Ligand-Based Virtual Screening / J.C. Baber, W.A. Shirley, Y. Gao, M. Feher. // Journal of Chemical Information and Modeling. – 2006. – T. 46, № 1. – C. 277-288.
256. *P.V. Karpov, I.I. Baskin, V.A. Palyulin, N.S. Zefirov*. Virtual screening based on one-class classification / P.V. Karpov, I.I. Baskin, V.A. Palyulin, N.S. Zefirov. // Doklady Chemistry. – 2011. – T. 437, № 2. – C. 107-111.
257. *P.V. Karpov, I.I. Baskin, N.I. Zhokhova, M.B. Nawrozkij, A.N. Zefirov, A.S. Yablokov, I.A. Novakov, N.S. Zefirov*. One-class approach: models for virtual screening of non-nucleoside HIV-1 reverse transcriptase inhibitors based on the concept of continuous molecular fields / P.V. Karpov, I.I. Baskin, N.I. Zhokhova, M.B. Nawrozkij, A.N. Zefirov, A.S. Yablokov, I.A. Novakov, N.S. Zefirov. // Russian Chemical Bulletin. – 2011. – T. 60, № 11. – C. 2418-2424.
258. *P.V. Karpov, I.I. Baskin, N.I. Zhokhova, N.S. Zefirov*. Method of continuous molecular fields in the one-class classification task / P.V. Karpov, I.I. Baskin, N.I. Zhokhova, N.S. Zefirov. // Doklady Chemistry. – 2011. – T. 440, № 2. – C. 263-265.
259. *P.V. Karpov, D.I. Osolodkin, I.I. Baskin, V.A. Palyulin, N.S. Zefirov*. One-class classification as a novel method of ligand-based virtual screening: The case of glycogen synthase kinase 3OI inhibitors / P.V. Karpov, D.I.

Osolodkin, I.I. Baskin, V.A. Palyulin, N.S. Zefirov. // Bioorganic & Medicinal Chemistry Letters. – 2011. – Т. 21, № 22. – С. 6728-6731.
260. *N.I. Zhokhova, I.I. Baskin*. Energy-based Neural Networks as a Tool for Harmony-based Virtual Screening / N.I. Zhokhova, I.I. Baskin. // Molecular Informatics. – 2017. – Т. 36, № 11. – С. 1700054.
261. *A. Bender*. Bayesian Methods in Virtual Screening and Chemical Biology // Chemoinformatics and Computational Chemical Biology / Bajorath J. – Totowa, NJ: Humana Press, 2011. – С. 175-196.
262. *S. Ai, Y. Bai, X. Liu*. Virtual Screening for COX-2 Inhibitors with Random Forest Algorithm and Feature Selection // Book Virtual Screening for COX-2 Inhibitors with Random Forest Algorithm and Feature Selection / Editor. – Barcelona, Spain: ACM, 2017. – С. 9-14.
263. *J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer*. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information / J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer. // Journal of Medicinal Chemistry. – 2005. – Т. 48, № 22. – С. 7049-7054.
264. *K.A. Carpenter, D.S. Cohen, J.T. Jarrell, X. Huang*. Deep learning and virtual drug screening / K.A. Carpenter, D.S. Cohen, J.T. Jarrell, X. Huang. // Future Medicinal Chemistry. – 2018. – Т. 10, № 21. – С. 2557-2567.
265. *R.N. Jorissen, M.K. Gilson*. Virtual Screening of Molecular Databases Using a Support Vector Machine / R.N. Jorissen, M.K. Gilson. // Journal of Chemical Information and Modeling. – 2005. – Т. 45, № 3. – С. 549-561.
266. *D. Plewczynski, S.A.H. Spieser, U. Koch*. Assessing Different Classification Methods for Virtual Screening / D. Plewczynski, S.A.H. Spieser, U. Koch. // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 3. – С. 1098-1106.
267. *R. Carbó, L. Leyda, M. Arnau*. How similar is a molecule to another? An electron density measure of similarity between two molecular structures / R. Carbó, L. Leyda, M. Arnau. // International Journal of Quantum Chemistry. – 1980. – Т. 17, № 6. – С. 1185-1189.
268. *M. Hahn*. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds / M. Hahn. // Journal of Chemical Information and Computer Sciences. – 1997. – Т. 37, № 1. – С. 80-86.
269. *J.A. Grant, B.T. Pickup*. A GAUSSIAN DESCRIPTION OF MOLECULAR SHAPE / J.A. Grant, B.T. Pickup. // Journal of Physical Chemistry. – 1995. – Т. 99, № 11. – С. 3503-3510.
270. *J.A. Grant, M.A. Gallardo, B.T. Pickup*. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape / J.A. Grant, M.A. Gallardo, B.T. Pickup. // Journal of Computational Chemistry. – 1996. – Т. 17, № 14. – С. 1653-1666.

271. *T.S. Rush, J.A. Grant, L. Mosyak, A. Nicholls*. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction / T.S. Rush, J.A. Grant, L. Mosyak, A. Nicholls. // Journal of Medicinal Chemistry. – 2005. – Т. 48, № 5. – С. 1489-1495.
272. *D.W. Ritchie, G.J.L. Kemp*. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces / D.W. Ritchie, G.J.L. Kemp. // Journal of Computational Chemistry. – 1999. – Т. 20, № 4. – С. 383-395.
273. *V.I. Pérez-Nueno, D.W. Ritchie, J.I. Borrell, J. Teixidó*. Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket / V.I. Pérez-Nueno, D.W. Ritchie, J.I. Borrell, J. Teixidó. // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 11. – С. 2146-2165.
274. *N.L. Max*. APPROXIMATING MOLECULAR-SURFACES BY SPHERICAL-HARMONICS / N.L. Max. // Journal of Molecular Graphics. – 1988. – Т. 6, № 4. – С. 210-210.
275. *N.L. Max, E.D. Getzoff*. SPHERICAL HARMONIC MOLECULAR-SURFACES / N.L. Max, E.D. Getzoff. // Ieee Computer Graphics and Applications. – 1988. – Т. 8, № 4. – С. 42-50.
276. *Q. Wang, K. Birod, C. Angioni, S. Grösch, T. Geppert, P. Schneider, M. Rupp, G. Schneider*. Spherical Harmonics Coefficients for Ligand-Based Virtual Screening of Cyclooxygenase Inhibitors / Q. Wang, K. Birod, C. Angioni, S. Grösch, T. Geppert, P. Schneider, M. Rupp, G. Schneider. // PLOS ONE. – 2011. – Т. 6, № 7. – С. e21554.
277. *V.I. Perez-Nueno, V. Venkatraman, L. Mavridis, T. Clark, D.W. Ritchie*. Using Spherical Harmonic Surface Property Representations for Ligand-Based Virtual Screening / V.I. Perez-Nueno, V. Venkatraman, L. Mavridis, T. Clark, D.W. Ritchie. // Molecular Informatics. – 2011. – Т. 30, № 2-3. – С. 151-159.
278. *W.S. Cai, X.G. Shao, B. Maigret*. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening / W.S. Cai, X.G. Shao, B. Maigret. // Journal of Molecular Graphics & Modelling. – 2002. – Т. 20, № 4. – С. 313-328.
279. *P.J. Ballester, W.G. Richards*. Ultrafast shape recognition to search compound databases for similar molecular shapes / P.J. Ballester, W.G. Richards. // Journal of Computational Chemistry. – 2007. – Т. 28, № 10. – С. 1711-1723.
280. *M.S. Armstrong, G.M. Morris, P.W. Finn, R. Sharma, W.G. Richards*. Molecular similarity including chirality / M.S. Armstrong, G.M. Morris,

P.W. Finn, R. Sharma, W.G. Richards. // Journal of Molecular Graphics & Modelling. – 2009. – T. 28, № 4. – C. 368-370.
281. *A.M. Schreyer, T. Blundell*. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints / A.M. Schreyer, T. Blundell. // Journal of Cheminformatics. – 2012. – T. 4.
282. *H. Li, K.-S. Leung, M.-H. Wong, P.J. Ballester*. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques / H. Li, K.-S. Leung, M.-H. Wong, P.J. Ballester. // Nucleic Acids Research. – 2016. – T. 44, № W1. – C. W436-W441.
283. *S.K. Kearsley, G.M. Smith*. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap / S.K. Kearsley, G.M. Smith. // Tetrahedron Computer Methodology. – 1990. – T. 3, № 6 PART C. – C. 615-633.
284. *D.J. Wild, P. Willett*. Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm / D.J. Wild, P. Willett. // Journal of Chemical Information and Computer Sciences. – 1996. – T. 36, № 2. – C. 159-167.
285. *S.K. Drayton, K. Edwards, N. Jewell, D.B. Turner, D.J. Wild, P. Willett, P.M. Wright, K. Simmons*. Similarity searching in files of three-dimensional chemical structures: Identification of bioactive molecules / S.K. Drayton, K. Edwards, N. Jewell, D.B. Turner, D.J. Wild, P. Willett, P.M. Wright, K. Simmons. // Internet Journal of Chemistry. – 1998. – T. 1, № 37. – C. CP3-U34.
286. *D.A. Thorner, D.J. Wild, P. Willett, P.M. Wright*. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials / D.A. Thorner, D.J. Wild, P. Willett, P.M. Wright. // Journal of Chemical Information and Computer Sciences. – 1996. – T. 36, № 4. – C. 900-908.
287. *D.A. Thorner, P. Willett, P.M. Wright, R. Taylor*. Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs / D.A. Thorner, P. Willett, P.M. Wright, R. Taylor. // Journal of Computer-Aided Molecular Design. – 1997. – T. 11, № 2. – C. 163-174.
288. *T. Cheeseright, M. Mackey, S. Rose, A. Vinter*. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation / T. Cheeseright, M. Mackey, S. Rose, A. Vinter. // Journal of Chemical Information and Modeling. – 2006. – T. 46, № 2. – C. 665-676.

СОДЕРЖАНИЕ


МОЛЕКУЛА
ПРЕДСТАВЛЕНИЕ
ХИМИЧЕСКИЙ
АТОМ
СТРУКТУРА
ДАННЫЕ
МОЛЕКУЛЯРНЫЙ
СОЕДИНЕНИЕ
РЕАКЦИЯ
СВЯЗЬ
СТРОКА
ГРАФ
ЧИСЛО
ПРОГРАММА
ПОВЕРХНОСТЬ
КОНФОРМАЦИЯ
ХЕМОИНФОРМАТИКА
ИНФОРМАЦИЯ
МЕТОД
ФРАГМЕНТ
ИСПОЛЬЗОВАНИЕ
МАТРИЦА
ПОИСК
НАЗВАНИЕ
ОПИСАНИЕ
ЗНАЧЕНИЕ
АЛГОРИТМ
СЛУЧАЙ
ТАБЛИЦА
SMILES