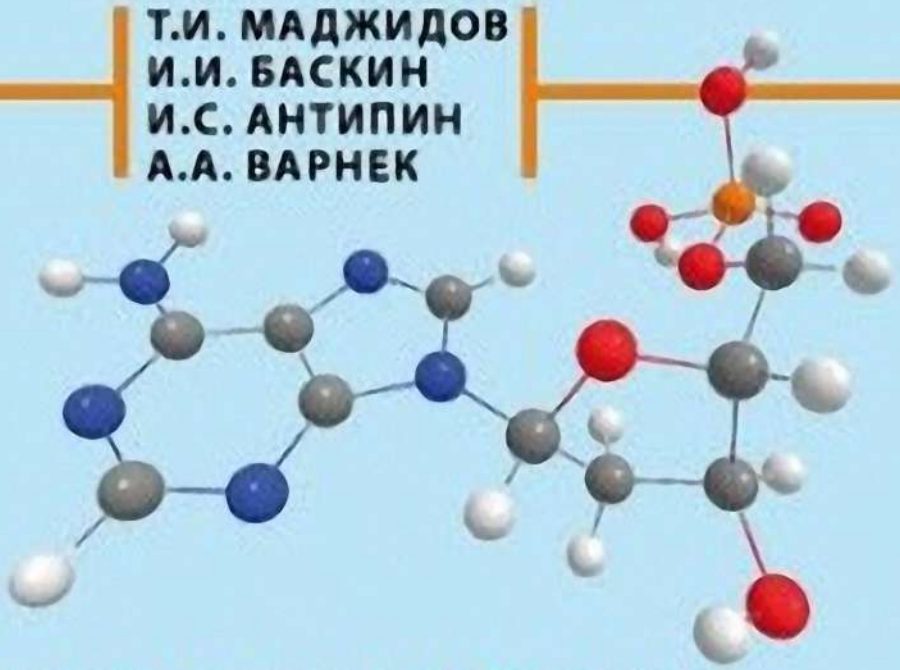
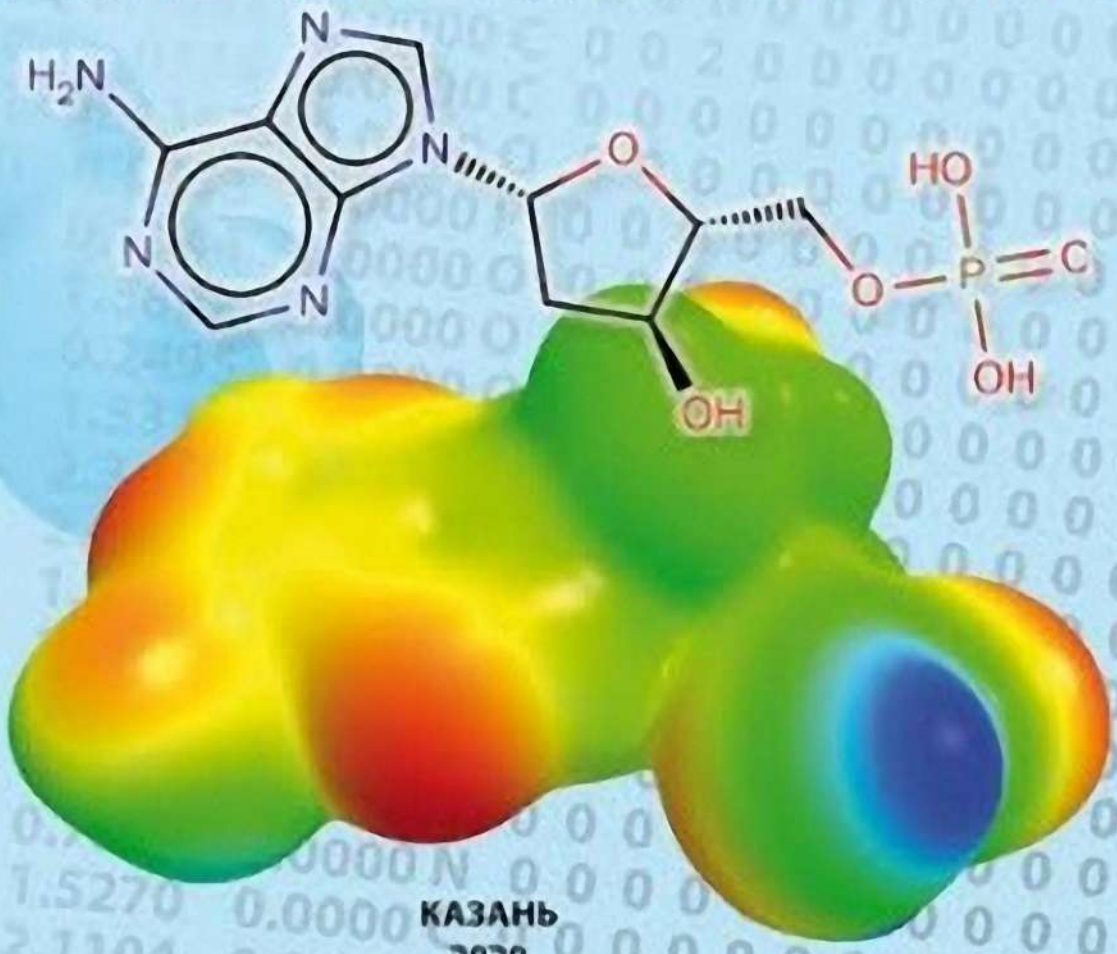
ВВЕДЕНИЕ В ХЕМОИНФОРМАТИКУ

Т.И. МАДЖИДОВ
И.И. БАСКИН
И.С. АНТИПИН
А.А. ВАРНЕК



КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКИХ СТРУКТУР



КАЗАНЬ
2020

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

Введение в хемоинформатику

Учебное пособие

Часть 1

Компьютерное представление химических структур

Казань
Москва
Страсбург
2020

Версия 1.0

Введение в хемоинформатику: Компьютерное представление химических структур: учеб. пособие / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. – Казань, Москва, Страсбург, 2020. – 176 с.

Данное пособие является первым из серии «Введение в хемоинформатику». Оно определяет предмет данной науки, а также дает подробную информацию о представлении химических объектов в виде графов, дескрипторов, молекулярных «отпечатков пальцев», широко распространенных в химии линейных нотаций SMILES, SLN, InChI и форматов файлов: MOL, SDF, MOL2, RDF и прочих.

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

Печатная версия пособия:
Введение в хемоинформатику: Компьютерное представление химических структур: учеб. пособие / Т.И. Маджидов, И.И. Баскин, И.С. Антипин, А.А. Варнек. – Казань: Казан. ун-т, 2013. – 174 с.

ПРЕДИСЛОВИЕ

Хемоинформатика – это мультидисциплинарное научное направление, возникшее на стыке химии, биологии, фармакологии, математики и информатики. Оно занимается обработкой накопленных экспериментальных данных о существующих химических элементах, а также развивает подходы, позволяющие заранее предсказывать химические, физические и биологические свойства новых, в том числе еще не синтезированных соединений.

В настоящее время элементы хемоинформатики преподаются во многих университетах мира, а в двух из них – Страсбургском (Франция) и Университете штата Индиана (США) – действуют магистерские программы по данной дисциплине. В России первая магистерская программа «Хемоинформатика и молекулярное моделирование» открыта в 2012 году в Казанском федеральном университете. Отдельные главы хемоинформатики преподаются в МФТИ и МГУ им. М.В. Ломоносова, и есть основания полагать, что данная дисциплина будет активно внедряться в учебные программы университетов нашей страны.

Задачей авторского коллектива стало создание современного учебника, последовательно описывающего основные концепции хемоинформатики и дающего минимальную, но достаточную информацию о теоретических методах и алгоритмах, на которые эта наука опирается. Принимая во внимание большой объем информации, авторы решили разбить учебник на несколько частей, публикуемых в виде отдельных пособий. Предполагается, что эти части будут опубликованы в 2013-2014 годах.

Предлагаемый учебник базируется на более чем десятилетнем опыте преподавания хемоинформатики в Страсбургском университете в рамках магистерской программы «Хемоинформатика и моделирование». В первой части учебника, названной «Компьютерное представление химических структур», читатель найдет подробную информацию о представлении химических объектов в виде графов, дескрипторов и молекулярных «отпечатков пальцев». Кроме того, показано, что химические объекты могут быть представлены и в виде широко распространенных в химии линейных нотаций SMILES, SLN, InChI, а также форматов файлов (MOL, SDF, MOL2, RDF и проч.).

При написании русскоязычного текста учебника авторы столкнулись с естественными проблемами перевода англоязычной терминоло-

гии, ибо для большинства устоявшихся в хемоинформатике понятий не существует общепринятых русских аналогий. Поэтому наряду с введенными новыми терминами приводятся их английские эквиваленты. Авторы надеются, что книга будет полезной не только для студентов и аспирантов, специализирующихся в области хемоинформатики, но и для широкого круга читателей – химиков, биологов, физиков, математиков.

Авторы выражают искреннюю благодарность д.х.н. В.П. Соловьеву, к.х.н. П.Г. Полищуку, д-ру В. Иленфельдту и к.х.н. А. Крутикову, взявших на себя труд прочтения текста книги, с последовавшими ценными замечаниями и советами, а также Т.И. Сибгатуллиной, аспирантам В.А. Афониной, А. Рахимбековой, А.А. Фатыховой, студентам Р. Пикалевой и К. Пикалевой за неоценимую помощь в подготовке иллюстративного материала.

кандидат химических наук,
доцент КФУ ***Т.И. Маджидов****,*
доктор физико-математических наук,
ведущий научный сотрудник
МГУ им. М.В. Ломоносова ***И.И. Баскин****,*
доктор химических наук, чл-корр. РАН,
профессор КФУ ***И.С. Антипин,***
доктор химических наук,
профессор Страсбургского университета (Франция) ***А.А. Варнек***

1. ХЕМОИНФОРМАТИКА КАК НАУЧНАЯ ДИСЦИПЛИНА

Хемоинформатика – относительно молодая дисциплина, объединившая несколько ранее независимых научных направлений:

– разработку компьютерных методов работы со структурной химической информацией, включая создание и оперирование химическими базами данных;

– моделирование связи между структурами химических соединений и их свойствами;

– компьютерное планирование синтеза химических соединений и предсказание путей химических превращений;

– автоматическую расшифровку структур химических соединений при помощи спектральных методов физико-химического анализа;

– молекулярный дизайн с использованием данных по структурам биологических мишеней (чаще всего белков).

В последние годы в результате интеграции с такими научными областями, как наука о материалах и нанотехнология, сформировалось и бурно развивается новое направление в хемоинформатике – информатика материалов.

Самую долгую историю из этих направлений имеет моделирование «структура – свойство»[1]. Пионерной работой в этом направлении следует считать создание Д.И. Менделеевым Периодической системы, позволившей предсказывать существование новых, еще не открытых химических элементов, и их свойства. Идеи о связи химической структуры (в ее современном понимании) с биологической активностью впервые были высказаны еще в конце XIX века. Современная методо-

[1] SAR/QSAR/QSPR – широко распространенное обозначение методов моделирования связи структуры вещества с его свойствами. SAR (англ. *structure-activity relationship*) – методы качественного предсказания и описания связи структуры с биологической активностью. QSAR (англ. *quantitative structure-activity relationship*) – методы количественного предсказания и описания связи структуры с биологической активностью. QSPR (англ. *quantitative structure-property relationship*) – методы количественного предсказания и описания связи структуры со свойствами молекул. Зачастую группу методов SAR/QSAR/QSPR называют просто QSAR.

логия SAR/QSAR/QSPR во многом обязана своим появлением трудам Л. Гаммета, К. Ганча, И. Гастайгера (вторая половина XX века).

Компьютерные базы данных, содержащие сведения о химических соединениях и реакциях, появились вместе с вычислительной техникой в середине XX века. Необходимость их создания была обусловлена накоплением настолько большого объема информации о химических объектах (веществах и реакциях), что стало сложно организовывать ее хранение в форме бумажных каталогов типа Бельштейна, Гмелина, Chemical Abstracts, а еще сложнее – осуществлять в них поиск требуемой информации. Естественным решением для возникших в связи с этим проблем стало активное внедрение компьютерной техники. Разработка первых химических баз данных осуществлялась главным образом специалистами, работающими в крупных компаниях: Chemical Abstract Service (Г. Морган и др.), SEAC (Л. Рэй и Р. Кирш), Du Pont (Д. Глак), Imperial Chemical Industries (проект CROSSBOW). Современные алгоритмы поиска информации в химических базах данных появились в значительной мере благодаря трудам двух ученых из Университета Шеффилда (Великобритания) – М. Линча и П. Виллета.

В 1958 году при помощи рентгеноструктурного анализа была впервые расшифрована структура биомолекулы. Накопление информации о пространственном строении белков и их комплексов с малыми молекулами, в сочетании с быстрым развитием вычислительной техники, привело к развитию методов, напрямую использующих информацию о биологической мишени при разработке новых биологически активных химических соединений. Это направление особенно интенсивно развивалось в фармацевтических компаниях, поскольку позволило существенно сократить расходы на разработку новых лекарств. Среди ученых, чьи работы внесли наиболее существенный вклад в данной области, следует упомянуть И. Кунтца и Г.-И. Бома.

По мере развития этих первоначально независимых научных направлений они стали все больше проникать друг в друга. Так, представление молекул в виде графов, набора фрагментов или битовых строк используется как в базах данных, так и в моделировании «структура – свойство». Сочетание построения моделей «структура – свойство», поиска в химических базах данных, а также молекулярного дизайна, основанного на структуре биомолекулы, привело к разработке методов виртуального скрининга, основанного на знании структур биологических макромолекул (англ. *structure-based virtual screening*).

В итоге, для обозначения возникшей в результате этой интеграции новой научной дисциплины, Фрэнк Браун в 1998 году ввел в своей работе [1] новый термин – «хемоинформатика».

Быстрому развитию хемоинформатики способствовал экспоненциальный рост числа известных химических соединений и реакций. В настоящее время, в наиболее представительной базе данных CAS REGISTRY, с 1907 по 2012 год зарегистрировано более 70 миллионов соединений, причем половина из них была синтезирована в течение последних 6 лет. Развитие экспериментальных методов высокопроизводительного скрининга привело к лавинообразному накоплению огромного количества данных. Так, созданная в 2004 году база PubChem содержит информацию о более чем 620 тыс. биологических испытаний, в каждом из которых было задействовано от нескольких десятков до нескольких десятков тысяч соединений. Для эффективной работы баз данных, обеспечивающих хранение, поиск и анализ химических соединений, реакций и соответствующей экспериментальной информации, хемоинформатика разрабатывает методы кодирования молекулярных структур и различные алгоритмы их поиска – по структуре, по подструктуре, по суперструктуре и подобию.

Наряду с большим объемом накопленных данных, все же ощущается их явная нехватка. Так, при поиске новых медикаментов для прогнозирования возможных побочных эффектов, необходимо оценить взаимодействие данной молекулы не только с выбранной биологической мишенью, но и с другими белками. Помимо этого, необходима информация о фармакокинетических параметрах, определяющих абсорбцию, распределение, метаболизм и выведение из организма, а также о токсичности. Принимая во внимание наличие не менее 5000 белков, являющихся потенциальными мишенями биологически активных молекул (англ. *druggable proteins*), экспериментальный скрининг всех существующих на сегодняшний день молекул представляется маловероятным. Отметим, что по самым скромным оценкам, число молекул, обладающих лекарственными свойствами, может достигать 10^{40}. Поэтому вполне очевидно, что теоретические подходы, позволяющие быстро и эффективно предсказывать свойства как реальных, так и гипотетических молекул, будут играть все более значимую роль в исследовательском процессе. Уже сегодня методы виртуального скрининга стали неотъемлемой частью процесса разработки лекарственных препаратов, и это не удивительно: разработка одного

лекарства стоит 1-1.5 миллиарда долларов, и хемоинформатика позволяет существенно уменьшить эти затраты.

Хотя на сегодняшний день хемоинформатика применяется, большей частью, в области фармацевтики, ее методы могут с успехом использоваться в любых других областях химии для поиска новых химических соединений и материалов, обладающих заданными свойствами. Например, дизайн ионных жидкостей (растворителей с уникальными свойствами) методами случайного перебора катионов и анионов непродуктивен из-за огромного числа (10^{18}!) возможных комбинаций этих компонентов. Направленный синтез материалов с заданными свойствами, такими как вязкость, электропроводность, плотность, температура плавления, становится возможным благодаря созданию предсказательных моделей, связывающих состав жидкости с ее свойствами. Известны работы по применению хемоинформатики для получения новых катализаторов, полимеров, веществ, избирательно связывающих конкретные металлы, а также для решения иных практических задач.

В последние годы методы хемоинформатики все активнее используются в области экологии и административного регулирования. Так, агенство по экологии США (U.S. Environment Protection Agency) применяет теоретически предсказанные параметры токсичности молекул в качестве важного дополнения к экспериментальным данным, представляемым производителями химикатов для принятия решения о допуске их на рынок. В Канаде и Дании также существуют государственные агентства, использующие модели QSAR для оценки влияния химических соединений на экологию.

В июне 2007 года в Европейском союзе вступила в силу Система регистрации, оценки, разрешения и ограничения химических веществ (т.н. REACH – Registration, Evaluation, Authorisation and Restrictions of Chemicals). В соответствии с ней, для веществ, произведенных или импортированных в количестве 1 тонны и более в год, производители и импортеры должны представить досье, содержащее несколько десятков параметров, описывающих физико-химические свойства соединения, его токсичность, мутагенность и другие. Для снижения стоимости экспериментальных испытаний предложено использовать расчетные методы. Для этого комиссия Организации экономического сотрудничества и развития разрабатывает правила для валидации моделей QSAR (т.н. «OECD principles for the Validation, for Regulatory Purposes, of QSAR Models»).

Таблица 1

Различные определения хемоинформатики
как области научных знаний

Ф. Браун [1]	1998	Использование информационных технологий и управления стали критической частью процесса разработки лекарственных препаратов. Хемоинформатика – это смешение информационных ресурсов для трансформации данных в информацию и информации в знания с целью предлагать правильные решения быстрее в области определения соединения-лидера.
Г. Пэрис [2]	1999	Хемоинформатика – это общий термин, включающий в себя дизайн, создание, организацию, управление, поиск, анализ, распространение, визуализацию и использование химической информации.
И. Гастайгер [3]	2003	Хемоинформатика – это применение методов информатики для решения химических проблем.
Ж.-Л. Фулон и А. Бендер [4]	2010	Хемоинформатика – это область, занимающаяся обработкой химической информации.
А. Варнек и И. Баскин [5]	2011	Хемоинформатика – это область теоретической химии, основанная на представлении молекул как объектов (графов или векторов) в химическом пространстве.

В настоящее время методы хемоинформатики широко применяются как в университетах и академических институтах, так и в научно-производственных предприятиях (особенно в области создания лекарственных препаратов). Хемоинформатика, как самостоятельная научная дисциплина, преподается в ряде университетов мира, по ней проводятся ежегодные конференции специалистов, издаются специализированные журналы. Ее становление прошло через ряд этапов, и это не могло не отразиться в характере формулировок ранних определений, данных разными авторами. Если вначале хемоинформатика ассоциировалась исключительно с разработкой лекарств (таблица 1), то более поздние работы показали, что эта область знаний может быть применима к моделированию любых химических, физических и биологических свойств соединений. Как и всякая другая дисциплина, она базируется на собственных концепциях и имеет явные отличия от комплементарных ей дисциплин.

Таблица 2

Взаимоотношения между различными ветвями теоретической химии

	Квантовая химия	**Молекулярное моделирование с использованием силовых полей**	**Хемоинформатика**
Молекулярная модель	Электроны и ядра	Атомы и связи (реже атомные ансамбли и связи между ними)	Графы и векторы дескрипторов
Механизм логического вывода	Дедуктивный >> Индуктивный	Дедуктивный ≈ Индуктивный	Дедуктивный << Индуктивный
Применение	Индивидуальные объекты (атомы, молекулы) или системы из небольшого числа объектов	Индивидуальные объекты или сложные системы, образованные большим числом объектов	Ансамбли объектов (для получения знаний или предсказания), индивидуальные объекты (для предсказаний). Объектами могут быть молекулы, комплексы, реакции, смеси, материалы, полимеры
Базовая концепция	Корпускулярно-волновой дуализм	Классическая механика	Химическое пространство
Базовые математические подходы	Уравнение Шредингера и приближенные способы его решения (уравнение Хартри-Фока, теория функционала плотности и др.)	Методы, основанные на использовании эмпирических силовых полей: молекулярная механика, динамика, методы Монте-Карло и возмущений свободной энергии	Математическая статистика, методы машинного обучения и интеллектуального анализа данных, теория графов

Одна из основных сфер применения методов хемоинформатики заключается в построении моделей, связывающих структуру и различные свойства химических соединений. Эта задача логично роднит хемоинформатику с двумя другими областями теоретической химии – квантовой химией и моделированием с использованием силовых полей. Эти три взаимодополняющих подхода отличаются используемыми молекулярными моделями, базовыми концепциями, механизмами логического вывода и типами рассматриваемых объектов (табл. 2). В отличие от молекулярных моделей, используемых в квантовой химии (электроны и ядра), и методов силовых полей (атомы и связи), в хемоинформатике молекулы обычно представлены при помощи молекулярных графов или связанных с ними векторов дескрипторов. Представление большинства химических объектов, в том числе реакций, смесей или материалов, обычно сводится к представлениям, близким тем, что используются для описания молекул. В хемоинформатике ансамбль графов или векторов, описывающих химические объекты, образует некоторое «химическое пространство», в котором должны быть определены соотношения между объектами. В отличие от реального физического пространства, химическое пространство не является уникальным: каждый ансамбль графов и дескрипторов определяет собственное химическое пространство. Переформулировав определение, данное в работе [5], можно определить **хемоинформатику как область знаний, основанную на представлении химических объектов как элементов химического пространства.**

Главными задачами современной хемоинформатики являются:

– организация эффективного хранения и поиска химической информации;

– разработка моделей, связывающих структуру химических объектов (молекул, смесей, реакций) и их свойства.

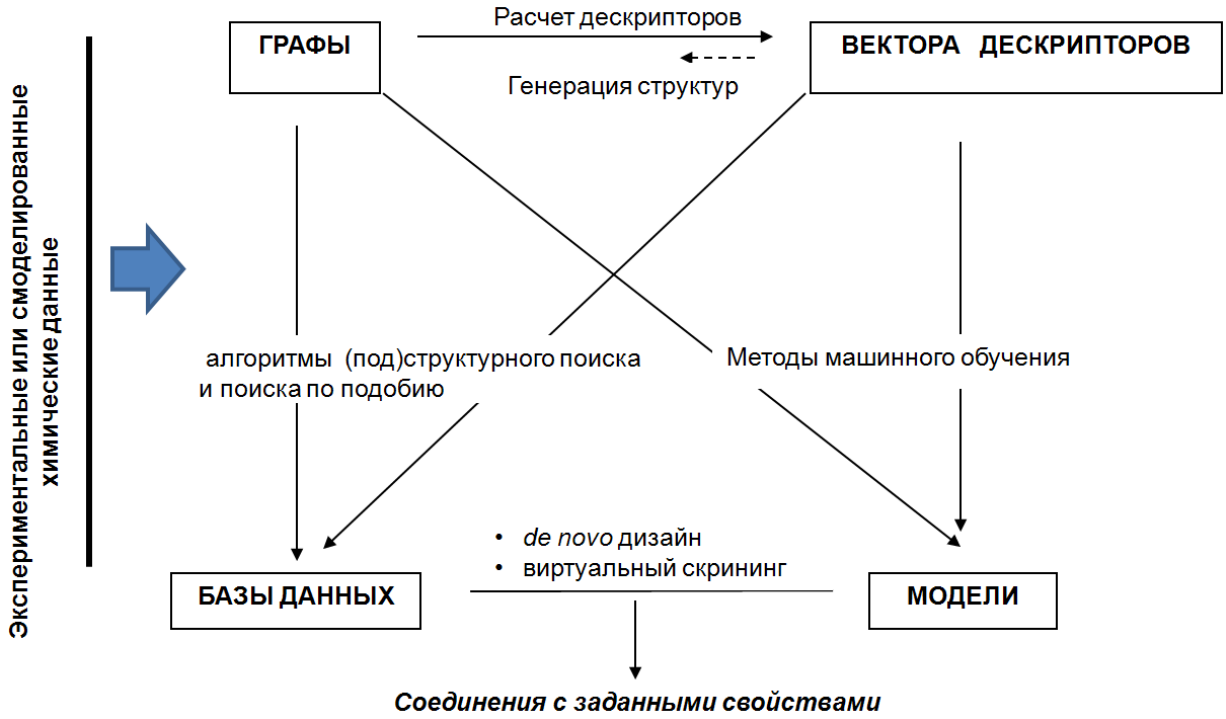


Рис. 1. Хемоинформатика от объектов
до главных областей применения

Рисунок 1 иллюстрирует связь между базовыми представлениями химических объектов, методами их обработки и двумя областями применения хемоинформатики. Задача поиска новых соединений с заданными свойствами, или *компьютерного дизайна химических соединений* (англ. *in silico design*), решается путем виртуального скрининга баз данных, содержащих информацию о реальных или гипотетических молекулах. Рисунок 1 схематически иллюстрирует применение методов хемоинформатики для поиска соединений с заданными свойствами. Экспериментальная информация о химических объектах поступает в базы данных, где она кодируется молекулярными графами, из которых генерируются векторы дескрипторов (структурные ключи, молекулярные отпечатки). Оба способа кодировки используются для организации и поиска структур в базе. Для построения моделей, как правило, используется выборка из базы данных. Модель может быть построена либо на молекулярных дескрипторах (как правило, отличных от применяемых для организации поиска в базах), либо непосредственно на графах. Компьютерный дизайн осуществляется с применением моделей ко всем объектам баз данных в процессе виртуального скрининга.

Данная простая схема, в принципе, может быть использована и для других приложений хемоинформатики. Так в *de novo* дизайне[2], база данных должна содержать молекулы, теоретически сгенерированные с учетом специально подобранных функциональных групп и соединяющих их линкеров. При автоматической расшифровке структур из спектров, база данных также содержит молекулы, сгенерированные компьютером. Модель рассчитывает спектр каждой из них, выбирая молекулу, для которой предсказанный спектр соответствует экспериментальному. При планировании синтеза модель применяется к одному соединению с целью отыскания цепочки химических превращений, ведущих к его образованию.

1.1. РАЗЛИЧИЕ И КОМПЛЕМЕНТАРНОСТЬ ХЕМОИНФОРМАТИКИ, КВАНТОВОЙ ХИМИИ И МЕТОДОВ СИЛОВЫХ ПОЛЕЙ

1.1.1. Базовая молекулярная модель

Общность и различие трех ветвей теоретической химии – хемоинформатики, квантовой химии и методов силовых полей – напрямую связаны с тем, как в каждой из них представлена химическая структура, то есть какова ее базовая молекулярная модель. Квантовая химия в явном виде рассматривает поведение ансамбля электронов и ядер, образующих молекулу путем решения уравнения Шредингера. Поскольку аналитическое решение этого уравнения возможно только для случая одноэлектронного атома, на практике используются приближенные методы решения, среди которых наиболее известны метод Хартри-Фока и методы теории функционала плотности. Так как такие расчеты требуют значительных вычислительных ресурсов, обычно они используются для моделирования одной молекулы или реакции, а также ансамбля из нескольких молекул. Расчеты, в основном, проводятся для газовой фазы, введение же поправок на растворитель часто требует использования эмпирически определяемых параметров и уход от чисто

[2] *De novo* дизайн лекарственных препаратов – это способ решения обратной задачи моделирования, в результате которой на основании знания структуры связывающего пакета биомолекулы создается наиболее эффективно взаимодействующее с ним химическое соединение.

неэмпирических (*ab initio*) методов в сторону частично эмпирических подходов.

Методы силовых полей основаны на «классическом» рассмотрении атомов и связей. Они используют эмпирические уравнения для аппроксимации потенциальной энергии молекулы как суммы вкладов от связей и несвязывающих взаимодействий.

Такой подход легко сопрягается с классической механикой, что позволяет вычислять молекулярные траектории (молекулярно-динамическое моделирование) путем численного решения классических уравнений движения Ньютона, равно как со статистической механикой для генерации больцмановского ансамбля (метод Монте-Карло) и с алгоритмами оптимизации нелинейных функций. Ввиду простоты математической модели и расчета потенциальной энергии, методы силовых полей могут применяться для изучения ансамблей из многих тысяч атомов, и поэтому легко использоваться для моделирования протеинов, растворов и других сложных систем.

Хемоинформатика рассматривает химическое соединение как единый *химический объект*. Обычно он представлен в виде графа или набора вычисляемых для него дескрипторов (вектор дескрипторов). Кроме химических соединений, объектами хемоинформатики являются также их смеси, материалы и химические реакции. Набор химических объектов образует *химическое пространство*, для которого определены взаимоотношения (например, отношения сходства) между этими объектами. Прогнозирование свойств новых химических объектов производится в хемоинформатике путем интерполяции свойств уже изученных объектов, что достигается путем построения моделей «структура – свойство» при помощи алгоритмов машинного обучения (математической статистики, интеллектуального анализа данных). Поскольку указанные вычисления выполняются очень быстро, полученные модели структура – свойство могут использоваться для виртуального скрининга больших баз данных. **Виртуальный скрининг** (от англ. *screening* – просеивание, обследование, отбор) **– это последовательное применение модели «структура – свойство» к собранным в специальной базе данных компьютерным представлениям химических объектов с целью определения тех объектов, которые обладают заданными характеристиками.** Для прогнозирования практически любого свойства, для которого имеется достаточно много экспериментальных данных, с помощью средств хемоинформатики

может быть построена модель «структура – свойство», что далеко не всегда возможно при использовании методов квантовой химии или силовых полей.

Таким образом, хемоинформатика, квантовая химия и молекулярное моделирование с использованием силовых полей являются различными, но взаимосвязанными областями. Действительно, методы квантовой химии привели к созданию популярных дескрипторов типа индексов электронного состояния, а методы молекулярной механики сделали возможным эффективное вычисление конформаций молекул, использующихся в различных приложениях хемоинформатики (QSAR, фармакофоры, докинг). С другой стороны, методы машинного обучения могут использоваться и используются для подстройки значений некоторых параметров квантово-химических и молекулярно-механических подходов.

Каждая из этих областей имеет свою сферу применимости, свои достоинства и недостатки. Хорошее знание всех трех частей теоретической химии необходимо для выбора наиболее подходящего инструмента решения конкретной научной проблемы.

1.1.2. Построение логического вывода

Одним из ключевых факторов, отличающих хемоинформатику от квантовой химии и методов силовых полей, является механизм построения *логического вывода* (англ. *inference*). Обычно выделяют два крайних случая механизма логического вывода – индуктивный и дедуктивный. Индуктивный подход подразумевает выявление общих закономерностей на основе большого числа отдельных наблюдений; дедуктивный, напротив, подразумевает возможность делать заключение о поведении системы на основании общей теоретической посылки. Иными словами, индуктивный вывод делается от частного к общему, а дедуктивный – от общего к частному.

Квантово-химическое моделирование представляет собой типичный пример дедуктивного вывода, когда на основе общей физической модели (квантовой механики) рассчитываются свойства конкретных химических объектов. Напротив, в хемоинформатике полагается, что мир слишком сложен для описания его ограниченным набором правил. Неполнота нашего знания приводит к изменению парадигмы логического вывода: вместо поиска точного решения, хемоинформатика использует аппарат статистического прогнозирования. Правила (модели)

хемоинформатики не выводятся исключительно из какой-либо строгой физической теории, а в значительной мере опираются на индуктивное обучение на основе имеющихся экспериментальных данных. В индуктивном обучении модели являются результатом выявления и обобщения характеристических особенностей (паттернов[3]) в данных о химических объектах. Более общие модели (сформулированные на большем числе данных) имеют больший шанс быть предсказательными. В теории статистического обучения, представляющей собой математическую основу хемоинформатики, предложено множество подходов для оценки обобщающей способности модели [6].

Индуктивное обучение в той или иной мере также используется в квантовой химии и методах силовых полей. Например, в случае квантовой химии параметризация функционала электронной плотности (псевдопотенциала) зачастую основана на подгонке получаемых результатов к экспериментальным данным. В полуэмпирических методах квантовой химии некоторые интегралы подобраны на основании тех или иных эмпирических данных. В случае методов силовых полей индуктивное обучение является почти настолько же важным, как и дедуктивное, поскольку расчет потенциальной энергии в них основан на большом числе эмпирически подбираемых параметров.

1.2. КОНЦЕПЦИЯ ХИМИЧЕСКОГО ПРОСТРАНСТВА

Как было указано К. Липинским и Э. Хопкинсом, химическое пространство можно рассматривать как аналог Вселенной, в которой положение звезд занимают химические соединения [7]. Любая попытка оценить количество химических соединений, которые могут быть потенциально синтезированы, приводит к астрономически большим числам порядка более 10^{60} [8] (оценено для органических молекул, содержащих до 30 атомов), что сравнимо с числом атомов во Вселенной,

---

[3] Слово «паттерн» – не очень распространенное заимствование от англ. *pattern* – образец, шаблон, модель, принцип, образ, форма, стереотип. В хемоинформатике оно используется для описания каких-либо характеристик данных (например, каких-то фрагментов молекул), которые важны для проявления тех или иных свойств. Таким образом, распознавание паттернов (в русскоязычной научной литературе более распространен термин «распознавание образов») является одним из основных элементов построения модели в хемоинформатике.

оцениваемое примерно 10^{80} [9]. При этом было показано [10], что число синтетически доступных и интересных в качестве лекарственных препаратов химических соединений много меньше – только 10^{36}. Очевидно, что даже это число настолько велико, что невозможно не только синтезировать все эти соединения, но и сгенерировать их структурные формулы при помощи всех компьютеров на Земле (их число оценивается между 1 и 2 млрд). Требуемое на такую генерацию время намного превысит время жизни Вселенной, равное примерно $4 \cdot 10^{17}$ секунд. Целью хемоинформатики является поиск рационального способа представления практически бесконечного химического пространства и выбор способа навигации в нем. Эффективные стратегии навигации в химическом пространстве критически важны для разработки новых биологически активных соединений и лекарственных препаратов. Основной причиной этого является неравномерное распределение биологически активных соединений в химическом пространстве, приведшее к образованию компактных регионов, подобным галактикам во Вселенной [7]. Это же справедливо и для других категорий химических соединений. Для изучения навигации в химическом пространстве существует даже специальный термин – *хемография*, что подчеркивает определенную аналогию с проблемами из области географии.

Хотя термин «химическое пространство» широко используется в литературе по хемоинформатике, он все еще четко не определен. Здесь мы даем возможное определение этому понятию: **химическое пространство – это набор химических объектов, для которых определено отношение, описывающие их сходство друг с другом**. Подобное отношение может быть задано, например, при помощи функции «расстояния» между объектами. В этом случае химическое пространство определяется двумя ключевыми характеристиками – способом представления химических объектов и функцией, на основе которой вычисляется расстояние между ними. Сейчас же остановимся на двух ключевых способах представления химических соединений: графах и векторах дескрипторов[4] (хотя их, на самом деле, больше).

---

[4] Вектор дескрипторов в данном случае обозначает столбец, заполненный значениями дескрипторов для данной молекулы. Использование слова «вектор» обусловлено тем, что данный столбец из N дескрипторов определяет вектор в N-мерном пространстве, направлен-

1.2.1. Представление химических объектов

Как уже было отмечено, в большинстве случаев представления химических объектов могут быть сведены к представлениям молекул. Например, реакции и смеси можно представить как набор молекул, реакции можно описать при помощи псевдо-молекул – конденсированных графов реакций, добавив новые типы «связей»: разрывающиеся, образующиеся и изменяющиеся. Кроме того, реакции и смеси химических соединений можно описать с использованием представлений, типичных для молекул, – векторов дескрипторов, матриц. Способы представления молекул и реакций подробно описаны ниже.

В настоящее время молекулы являются основным объектом химического пространства. В хемоинформатике молекулы рассматриваются как информационные объекты, обладающие определенной структурой и свойствами. При помощи специальных наборов меток для вершин и ребер графов можно специфицировать различные типы атомов и связей. Метка вершины молекулярного графа (например, символ химического элемента) идентифицирует тип атома, а метка ребра – тип связи. Метка связи может быть как обычным порядком связи, так и обозначать координационные, водородные связи, или динамические (разрывающиеся, образующиеся и изменяющиеся в ходе реакции) типы связей. Последний способ позволяет кодировать химические реакции. Более сложные химические системы, такие как полимеры и смеси, могут быть представлены с помощью наборов графов.

Для определенных целей используются более обобщенные представления химических структур. К примеру, в фармакофорном анализе вершины графа могут обозначать фармакофорные центры (H-доноры, H-акцепторы, катионный, анионный центры, алифатические и ароматические атомы), тогда как относительное расположение центров может быть задано с помощью топологических или геометрических расстояний. В обобщенных структурах (т.н. структурах Маркуша) одна вершина графа может соответствовать нескольким атомам или целым подструктурам (например, заместителям).

Другой важный способ представления химических структур основан на использовании молекулярных дескрипторов. В соответствии с

ный из начала координат к точке, представляющей объект, закодированный данным набором дескрипторов.

определением, данным Р. Тодескини и В. Консонни [11] (здесь мы приводим несколько сокращенный вариант определения), **дескриптор – это числовой результат некоторого стандартизованного эксперимента, либо финальный результат логической и математической процедуры, которая однозначно трансформирует структурную информацию о химическом объекте в число**. В качестве дескрипторов могут использоваться как экспериментально определенные характеристики химических объектов (например, экспериментально найденный коэффициент распределения вода-октанол), так и теоретически определяемые характеристики (например, число фенильных групп в молекуле, ВЗМО, рассчитанный коэффициент вода-октанол).

Дескрипторное представление химических объектов чрезвычайно распространено в хемоинформатике, поскольку:

– множество различных дескрипторов может быть рассчитано из одного и того же молекулярного графа;

– они инвариантны по отношению к перенумерации атомов в молекуле;

– модели «структура-свойство», построенные на дескрипторах с понятным структурным либо физико-химическим смыслом, легко могут быть интерпретированы;

– с использованием дескрипторов может проводиться индуктивный перенос знаний, полученных в рамках одной модели, на другую модель;

– химическое пространство, построенное на дескрипторном представлении химических объектов, является обычным Эвклидовым, что упрощает с ним работу по сравнению с математически значительно более сложными типами пространств, которые можно построить непосредственно на графах.

Дескрипторы могут быть рассчитаны не только для отдельных молекул, но и для реакций, смесей, материалов и даже нано-частиц. В настоящее время существует более 5 000 различных дескрипторов. Кроме того, число фрагментных дескрипторов практически не ограниченно. Дескрипторы используются для организации поиска в химических базах данных (на этапе скрининга), для создания моделей, связывающих структуру и свойства молекул (SAR/QSAR/QSPR моделей), в поиске по сходству, кластеризации соединений и для множества других целей. Различные типы дескрипторов химических структур будут подробно изложены в отдельном пособии этого учебника.

1.2.2. Соотношение между объектами в химическом пространстве

Химическое сходство (молекулярное или химическое подобие) является мерой соотношения между объектами в химическом пространстве. Это одна из основных концепций хемоинформатики, широко используемая в виртуальном скрининге и компьютерном дизайне новых соединений. *Принцип сходства* (англ. *similarity principle*) можно сформулировать следующим образом: **структурно сходные химические объекты с большей вероятностью обладают близкими свойствами, чем несходные (различные)**. С точки зрения применений этого принципа к классификационным проблемам – это значит, что соединения, относящиеся к одному классу (например, обладающие определенным типом биологической активности), обычно располагаются близко друг к другу в компактных областях химического пространства. При решении регрессионной задачи это означает, что кривая, описывающая зависимость изучаемого свойства от характеристик структуры, должна быть как можно более гладкой (без резких подъемов или спусков). Однако очевидно, что сходство между объектами зависит не только от метрики (способа вычисления расстояния), но также от выбранных дескрипторов или способа сопоставления графов.

Меры сходства между химическими объектами могут быть рассчитаны с использованием различных типов представления химических структур, в частности, молекулярных графов, векторов дескрипторов, молекулярных полей (например, функции электронной плотности) и некоторых других. Важным видом мер подобия химических объектов являются *ядра сходства* (англ. *kernels*), которые широко используются в задачах моделирования «структура – свойство».

Исторически первыми способами расчета мер сходства молекулярных графов без промежуточного вычисления молекулярных дескрипторов стали методы, основанные на поиске максимальных общих подграфов (MCS – от англ. *Maximum Common Subgraph*) для пар молекул. Недостатком данного подхода является высокая вычислительная сложность даже при помощи современных алгоритмов (например, при помощи поиска клик графов совместимости по алгоритму Брона-Кербоша [12]). Другой способ оценки мер сходства графов, не требующий вычисления молекулярных дескрипторов, основан на использовании т.н. *графовых ядер сходства* (англ. *graph kernels*) [13]. Графовые ядра сходства позволяют в неявном виде проецировать химическое

пространство, построенное на молекулярных графах, в некоторое векторное пространство, называемое *пространством признаков* (англ. *feature space*). Это позволяет работать с химическим пространством, определенным на молекулярных графах, практически с такой же легкостью, как с пространством, построенным с использованием молекулярных дескрипторов.

Наиболее широко используемые в хемоинформатике меры сходства молекул основаны на векторах дескрипторов. В этом случае значения дескрипторов определяют координаты точки, соответствующей химическому объекту, в пространстве, осями которого являются дескрипторы. Расстояния между точками задают отношение близости соответствующих объектов. В качестве таких расстояний могут использоваться различные типы метрик – евклидово, манхэттенское расстояние, расстояние Махаланобиса, Минковского и др. Также могут использоваться разнообразные индексы сходства между объектами – коэффициент Танимото, Дайса, Карбо (косинуса), Тверского.

1.3. ХЕМОИНФОРМАТИКА И СВЯЗАННЫЕ С НЕЙ ДИСЦИПЛИНЫ

Остановимся на сходстве и отличии хемоинформатики от сходных научных дисциплин.

1.3.1. Хемоинформатика и хемометрика

Д. Массарт [14] определил **хемометрику как химическую дисциплину, которая использует математику, статистику и формальную логику:**

– для дизайна и выбора оптимальной экспериментальной процедуры;

– обеспечения максимально обоснованной химической информацией на основании анализа химических данных;

– получения знаний о химических системах.

В основном, для хемометрики совершенно не важна информация о химической структуре, следовательно, она пересекается с хемоинформатикой лишь постольку, поскольку обе дисциплины используют методы машинного обучения для решения химических проблем. Хемометрика широко используется в экспериментальном дизайне, химической инженерии, аналитической химии и изучении спектров – обла-

стей, в которых необходимо полноценное изучение и использование многопараметрических данных.

1.3.2. Хемоинформатика и биоинформатика

В отличие от хемоинформатики, имеющей дело с молекулами «химического размера», биоинформатика использует вычислительные инструменты для изучения структуры и функций биомолекул (протеины, нуклеиновые кислоты). Это очень большая область, включающая 3D (использование методов силовых полей и квантовой механики) и 1D моделирование (выравнивание и наложение последовательностей). В последнем случае биомолекула представляется как строка символов, обозначающих строительные блоки биомолекул (нуклеиновых оснований или аминокислот). Графы и векторы значений фиксированного размера используются широко в хемоинформатике и крайне редко – в биоинформатике. В данном смысле хемо- и биоинформатика являются взаимодополняющими науками.

С другой стороны, хемоинформатика и биоинформатика взаимопроникают друг в друга. Основной областью взаимопроникновения этих двух дисциплин является дизайн лекарственных препаратов. Методы моделирования и виртуального скрининга, основанные на структуре биологической мишени, а также *de novo* дизайна лекарственных препаратов, требуют знания трехмерной структуры биомолекул, которые могут быть созданы с использованием средств биоинформатики. В то же время оценка эффективности взаимодействия лиганда (в хемо- и биоинформатике словом «лиганд» обозначается малая молекула, в супрамолекулярной химии называемая также «гостем») с рецептором (биомолекулой, «хозяином») производится с использованием векторного представления молекул. Такие средства, как докинг и скоринг, фармакофорный поиск, основанные на структуре биомолекулы, относятся к данному типу. Таким образом, хемоинформатика – большая область, включающая в себя множество различных аспектов и задач.

1.4. ЛИТЕРАТУРА

Основные журналы в области хемоинформатики:

- Journal of Chemical Information and Modeling (бывш. Journal of Chemical Documentation), *Изд-во Американского химического общества*, http://pubs.acs.org/journal/jcisd8

- Molecular Informatics (бывш. QSAR & Combinatorial Science, Quantitative Structure-Activity Relationships), *Изд-во Wiley*, http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291868-1751
- Journal of Cheminformatics, *Изд-во Chemistry Central*, http://www.jcheminf.com/
- Journal of Medicinal Chemistry, *Изд-во Американского химического общества*, http://pubs.acs.org/journal/jmcmar
- Journal of Computer-Aided Molecular Design, *Изд-во Springer*, http://link.springer.com/journal/10822
- SAR & QSAR in Environmental Research, *Изд-во Taylor & Francis*, http://www.tandfonline.com/toc/gsar20/current
- Bioorganic & Medicinal Chemistry, *Изд-во Elsevier*, http://www.journals.elsevier.com/bioorganic-and-medicinal-chemistry/
- Molecular Diversity, *Изд-во Springer*, http://www.springer.com/-ife+sciences/biochemistry+%26+biophysics/journal/11030
- Combinatorial Chemistry & High Throughput Screening, *Изд-во Bentham Science*, http://www.benthamscience.com/cchts/
- International Journal of Chemoinformatics and Chemical Engineering, *Изд-во IGI Global*, http://www.igi-global.com/journal/international-journal-chemoinformatics-chemical-engineering/1176

Избранные книги по хемоинформатике:

Учебные:

- *Leach A.R.* An Introduction to Chemoinformatics / A.R. Leach, V.J. Gillet. – Dortrecht: Kluwer Academic Publishers, 2003. – 259 p.
- *Gasteiger J.* Chemoinformatics. A textbook / J. Gasteiger, T. Engel. – Weinheim: John Wiley & Sons, 2003. – 680 p.

Научные:

- *Gasteiger J.* Handbook of Chemoinformatics: From Data to Knowledge / J. Gasteiger, T. Engel. – Weinheim: Wiley-VCH, 2003. – 1870 p.

- *Todeschini R.* Handbook of Molecular Descriptors / R. Todeschini, V. Consonni. – Weinheim: Wiley-VCH, 2009. – 2nd Edition. – 967 p. (V.1), 252 p. (V. 2).
- *Varnek A.* Chemoinformatics approach to Virtual Screening / A. Varnek, A. Tropsha. – Cambridge: Royal Society of Chemistry, 2008. – 338 p.
- *Oprea T.I.* Chemoinformatics in Drug Discovery / T.I. Oprea. – Weinheim: Wiley-VCH, 2005. – 493 p.
- *Sotriffer C.* Virtual Screening. Principles, Challenges, and Practical Guidelines / C. Sotriffer. – Weinheim: Wiley-VCH, 2011. – 519 p.
- *Faulon J.-L.* Handbook of Chemoinformatics Algorithms / J.-L. Faulon, A. Bender. – Boca Raton: CRC Press, 2010. – 440 p.

1.5. КОНФЕРЕНЦИИ

Регулярные международные конференции по хемоинформатике

- Секция «Chemical Informatics» на ACS Meeting, США (ежегодная).
- EuroQSAR, Вена, Австрия (раз в 2 года).
- International Conference on Chemical Structures, Нидерланды (раз в 2 года).
- Sheffield Conference on Chemoinformatics, Шеффилд, Великобритания (раз в 3 года).
- German Conference on Chemoinformatics, Гослар, Германия (ежегодная)
- Strasbourg Summer School on Chemoinformatics, Страсбург, Франция (раз в 2 года)

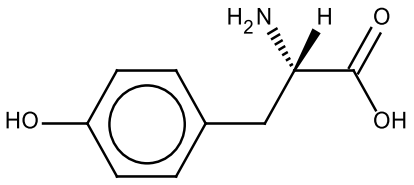
2. ПРЕДСТАВЛЕНИЕ МОЛЕКУЛ

Химия – это наука о веществах, их строении, свойствах и взаимопревращениях, происходящих в ходе химических реакций. Все вещества, как известно, состоят из атомов, которые путем образования химических связей способны связываться друг с другом, образуя молекулы. Согласно определению, которое было дано на международном съезде химиков в г. Карлсруэ (Германия) в 1860 г., **молекула – наименьшая частица химического вещества, обладающая всеми его химическими свойствами**. Как следует из этого определения, для однозначной идентификации индивидуального химического вещества достаточно описать строение любой из образующих его идентичных молекул. Именно поэтому молекулы являются центральными объектами исследования в химии и, как следствие, в хемоинформатике.

За последующие полтора столетия представление о строении молекул неоднократно уточнялось в соответствии с очередными достижениями в понимании физической природы химических веществ. Согласно недавнему определению, содержащемуся в «Золотой книге ИЮПАК» [15], **молекула – это электрически нейтральная частица, состоящая, по крайней мере, из двух атомов, которой на поверхности потенциальной энергии соответствует углубление, способное вместить хотя бы одно колебательное состояние**. Вне всякого сомнения, последнее определение является более строгим с точки зрения физической и квантовой химии. С его помощью можно, например, провести четкий водораздел между понятием молекулы и понятиями атома, иона и переходного состояния химической реакции. Тем не менее, с точки зрения хемоинформатики, практическую пользу представляет самое первоначальное понятие о молекуле как о наименьшей частице, обладающей всеми свойствами химического вещества. Подобная широкая трактовка позволяет, отвлекаясь от излишних физических деталей, использовать различные *представления молекул* для идентификации соответствующих химических веществ в электронных базах данных, создание и организация работы с которыми является одной из ключевых задач хемоинформатики. В ходе дальнейшего изложения термины «молекула», «химическое соединение» и «химическая структура» будут употребляться как синонимы.

Под представлением молекулы понимается любой способ ее описания. В качестве примеров различных представлений можно при-

вести структурную формулу соединения, в которой атомы (кроме углерода и водорода) изображаются символами элементов, а связи – черточками, соединяющими атомы (рис. 2 а). По начертанию связей в структурной формуле можно определить их порядок, принадлежность к ароматическим циклическим системам, а также стереохимические особенности молекулы (в частности, хиральные центры). Таким образом, химик может понять строение молекулы и примерное расположение атомов в пространстве. Это представление позволяет судить не только о расположении атомов, но и об электронном строении молекул. Другими представлениями, широко используемыми в химии, являются брутто-формула химического соединения (рис. 2 б), его тривиальное название (рис. 2 в) и название в соответствии с номенклатурой ИЮПАК (рис. 2 г). Брутто-формула вещества отражает стехиометрический состав молекул и в каталогах обычно записывается в соответствии с *системой Хилла*, в которой первыми указываются атомы углерода и водорода, а остальные атомы записываются в алфавитном порядке (рис. 2 б).



а)

$C_9H_{11}NO_3$

б)

L-тирозин

в)

(2S)-2-амино-3-(4-гидроксифенил)-пропановая кислота

г)

Рис. 2. Различные химические представления молекулы L-тирозина:
а) структурная формула, б) брутто-формула, в) тривиальное название,
г) название в соответствии с правилами ИЮПАК

2.1. ОСОБЕННОСТИ ПРЕДСТАВЛЕНИЯ МОЛЕКУЛ В ХЕМОИНФОРМАТИКЕ

Становление хемоинформатики как науки связано с накоплением огромного количества экспериментальных данных и необходимостью ими оперировать: хранить информацию о химических соединениях и реакциях, осуществлять ее поиск и извлечение. Первоначально для этого использовались напечатанные на бумаге каталоги соединений (наиболее известны в России – справочники Бельштейна (Германия), Chemical Abstracts (США) и Реферативный журнал «Химия» (Россия)), поиск в которых осуществлялся вручную по стехиометрическому со-

ставу (химической формуле) и названию химического соединения. Развитие компьютерной техники и накопление огромного количества информации привело к пониманию того, что эти задачи могут эффективно решаться только с использованием компьютеров и технологии компьютерных баз данных.

2.1.1. Требования к кодирующим представлениям молекул

Очевидно, что не любое представление молекулы является *кодирующим*, т.е. способным быть использованным для кодирования, хранения, поиска и извлечения информации о химических структурах при помощи компьютера. Можно выделить несколько требований, которым должны удовлетворять кодирующие представления молекул:

1. Легкость обработки при помощи компьютера. Например, графическое изображение (фотография, рисунок) структурной формулы химического соединения понятно химику, однако оно крайне сложно в обработке при использовании компьютеров, и поэтому не является кодирующим.

2. Высокая емкость. Для использования при хранении и оперировании информацией оно должно занимать как можно меньше места в памяти. Поэтому избыточность информации является нежелательной.

3. Эффективность. Желательно, чтобы для работы с кодирующими представлениями могли применяться высокоэффективные алгоритмы обработки информации.

4. Уникальность. Желательно, чтобы одной молекуле могло соответствовать только одно ее представление. Отметим, что это, казалось бы, очевидное требование часто бывает сложно удовлетворить на практике. Целый ряд представлений (например, при помощи матрицы смежности молекулярного графа, см. ниже) зависит от первоначально выбранной нумерации атомов в молекуле. Процесс выбора уникального представления из множества возможных вариантов называется канонизацией.

5. Однозначность. Каждому представлению в идеальном случае должна соответствовать только одна молекула. Этому требованию не удовлетворяет, например, брутто-формула из-за возможности существования множества структурных изомеров.

К сожалению, единого универсального представления молекул, идеально удовлетворяющего всем этим признакам, не существует. Вследствие этого выбор того или иного кодирующего представления

для использования в конкретной информационной системе определяется поставленными при ее разработке задачами.

2.1.2. Виды представлений

Представления молекул базируются на том или ином уровне их рассмотрения. В соответствии с этим их можно классифицировать на:

☐ 1D-представления. Они зависят только от стехиометрического состава молекулы. Примером является брутто-формула.

☐ 2D-представления, содержащие информацию о молекулярной связности, т.е. о том, как атомы в молекуле связаны друг с другом посредством химических связей. Примером является структурная формула химического соединения. Математическим объектом, лежащим в основе 2D-представлений, является молекулярный граф[5].

☐ 3D-представления, которые зависят от положения всех атомов в пространстве. Данный вид представления определяет конформацию молекулы.

☐ 4D-представления[6], описывающие гибкость молекул. Как правило, они содержат информацию о множестве конформаций, которые может принимать молекула. Типичным примером служат т.н. «молекулярно-динамические траектории», показывающие положения каждого атома в пространстве в разные моменты времени.

Функциональные особенности и особенности применения представлений зависят, однако, не только от их размерности, но также от того, каким образом данное представление организовано. По организации представлений можно выделить:

- линейное представление – состоит из буквенно-цифровой строки;

---

[5] Обозначение 2D в данном случае отнюдь не свидетельствует о планарности молекул. Оно традиционно используется для обозначения молекулярной связности в связи с тем, что большинство молекулярных графов являются планарными, т.е. узлы (соответствующие атомам) можно расположить на плоскости таким образом, чтобы ребра (соответствующие химическим связям) не пересекали друг друга.

[6] Обозначение 4D связано с тем, что в рамках 4D-представления молекулы положение каждого атома может быть описано четырьмя координатами – тремя пространственными и временем (если рассматривается молекулярно-динамическая траектория)

- представления молекулярных графов – кодируют молекулярную связность с использованием чисел, организованных в строку, матрицу или таблицу;
- трехмерные представления – кодируют трехмерную структуру молекулы с использованием таблиц, задающих либо пространственные координаты атомов (в декартовой либо косоугольной системе координат), либо набор внутренних координат (длин связей, валентных и торсионных углов);
- четырехмерные представления – кодируют набор различных трехмерных структур молекул при помощи многомерных таблиц. Кроме того, они могут содержать дополнительную информацию, описывающую индивидуальные конформации: энергию, момент времени (для молекулярно-динамической траектории).

2.2. ЛИНЕЙНЫЕ ПРЕДСТАВЛЕНИЯ

Линейные представления молекул – это буквенно-цифровые строки, составленные по тем или иным правилам. По сути, они представляют собой перевод структурной формулы молекулы (а в некоторых случаях реакций и субструктур) в строку, части которой отражают те или иные структурные особенности соединения.

2.2.1. Химическая номенклатура как линейное представление

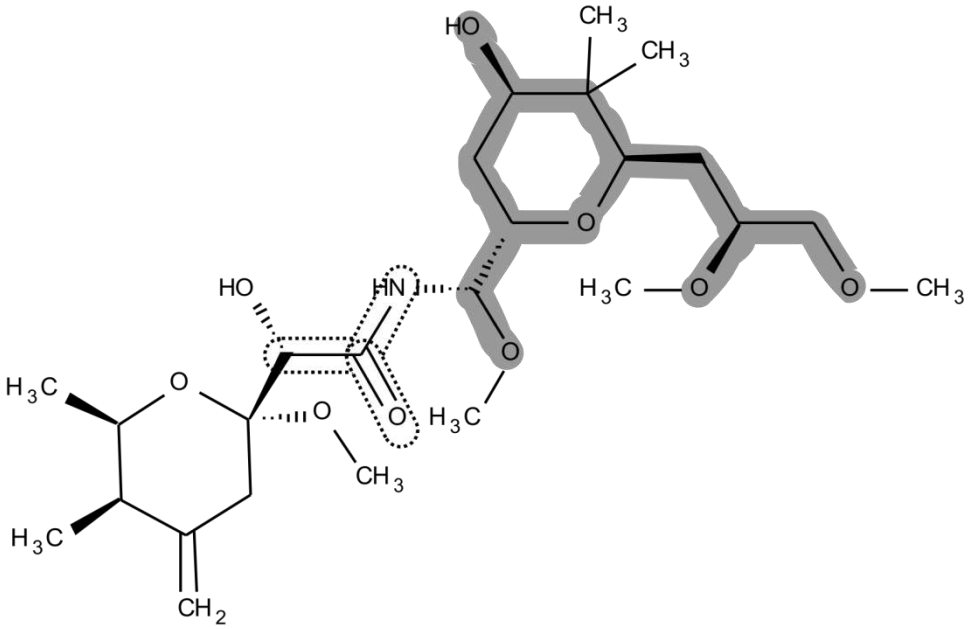
Название химического соединения является традиционным способом его линейного представления, появившимся еще в докомпьютерную эру. В химии достаточно широко используются три вида названий (имен) – тривиальное, рациональное и систематическое. Совокупность правил построения названий каждого вида образует соответствующую номенклатуру.

Тривиальные названия химических соединений обычно коротки и легки для запоминания. Как правило, они отражают свойства химических соединений (например, фосфор означает «светящийся»), происхождение или способ получения (например, салициловая кислота – от «*salix*», ива), а также их структуру (например, краун – от «*crown*», корона; каликсарен – от «*calix*», ваза). К этой группе имен можно отнести также торговые и устоявшиеся имена соединений (аспирин, диклофенак, витамин С). Наборы тривиальных названий химических со-

единений содержатся как в многочисленных бумажных справочниках (каталогах), так и в электронных базах данных, поиск в которых дает возможность восстановить структуры соответствующих молекул. Первым существенным недостатком тривиальных названий является отсутствие процедуры их автоматического формирования для новых соединений без участия химика. Вторым их существенным недостатком является принципиальная невозможность восстановить структуру молекулы из тривиального названия, отсутствующего в доступных каталогах. Поэтому они практически не используются как основной способ представления молекул в хемоинформатике, однако в специфических базах данных может присутствовать как дополнительный вид представления. В силу их широкой распространенности, тривиальные имена зачастую могут использоваться как часть систематических имен. Так, например, в соответствие с систематической номенклатурой, правильным будет имя «2-аминофенол», а не «2-амино-1-гидроксибензол» и не «2-амино-1-гидроксициклогексатриен».

Возникновение и развитие рациональной[7] и сменившей ее впоследствии систематической номенклатуры было связано с необходимостью создания общепринятых правил для формирования имен растущего числа соединений. Первые правила систематической номенклатуры были приняты в 1892 году на Международной конференции по реформированию номенклатуры в Женеве и известны как «Женевская номенклатура». Образованная в 1922 году ИЮПАКом Комиссия по номенклатуре химических соединений продолжает развитие номенклатуры органических соединений, взяв за основу «Женевскую номенклатуру». Работа комиссии продолжается до настоящего времени [16-19]. Правила систематической номенклатуры сложны даже для специалиста и допускают определенные разночтения (рис. 3). Они не обязательно дают уникальное название для каждого соединения (т.е. могут существовать разные названия одного вещества, не противоречащие правилам ИЮПАК), но эти названия всегда будут однозначными (одному названию будет соответствовать только одно вещество).

[7] Поскольку рациональная номенклатура крайне устарела, она упоминается здесь лишь как одна из первых попыток систематизации, мы на ней подробно не останавливаемся.



(2S)-N-[(S)-[(2R,4S,6S)-6-[(2S)-2,3-диметоксипропил]-4-гидрокси-5,5-диметилоксан-2-ил](метокси)метил]-2-гидрокси-2-[(2R,5R,6R)-2-метокси-5,6-диметил-4-метилиденоксан-2-ил]ацетамид

2,6-ангидро-3,5,7-тридеокси-1-C-[[(2S)-гидрокси[(2R,5R,6R)-тетрагидро-2-метокси-5,6-диметил-4-метилен-2H-пиран-2-ил]ацетил]амино]-5,5-диметил-1,8,9-три-O-метил-D-манно-нонитол

Рис. 3. Различные систематические имена для одного соединения. Выделена родительская структура при построении каждого названия

Наименование химического соединения в соответствии с систематической номенклатурой всегда содержит название основного (родительского) фрагмента, который может быть как одним атомом (редкий случай), так и цепочкой атомов, циклом или каркасной структурой. Заместители при атомах родительской структуры, которые ведут к получению искомой молекулы, обозначаются в названии с использованием суффиксов и префиксов (приставок)[8]. Положение заместителей указывается с использованием нумерации положений родительской структуры в соответствии с правилами старшинства заместителей[9].

[8] В номенклатуре ИЮПАК указывается также, что инфиксы (вставки внутри корня) являются одним из способов образования имени. Однако существование инфиксации в русском языке подвергается сомнению.

[9] Краткое описание правил ИЮПАК на русском языке можно найти в книге: *Газизов М.Б., Хайруллин Р.А., Каримова Р.Ф.* Номенкла-

Таблица 3

Преимущества и недостатки использования номенклатурных названий химических соединений в качестве линейных представлений

Преимущества	Недостатки
Тривиальные (и торговые) имена	
✓ Короткие, звучные, легко запоминаются ✓ Широко распространенные ✓ Однозначные ✓ Существуют не только для молекул, но и реакций, материалов, смесей, веществ (царская водка, брожение и т.п.)	○ Существует очень много ○ Из названия невозможно восстановить структуру ○ Не понятно, каким образом формируется название стереоизомеров ○ Неуникальные
Систематическая номенклатура (IUPAC)	
✓ Стандартизованные способы создания ✓ Определены способы формирования имени стереомеров ✓ Широко распространенные ✓ Однозначные ✓ Позволяют восстанавливать структуру из названия ✓ Существуют уникальные «предпочтительные» систематические имена	○ Сложные правила номенклатуры, занимающие несколько книг ○ Допустимы альтернативные «общие» систематические имена ○ Сложные, трудно запоминаемые и не звучные имена ○ Длинные (включают много символов) ○ Только для молекул

Изомерия соединений при двойной связи указывается с использованием приставок цис- или транс-, либо Z- или E-, что означает, что старшие заместители находятся по одну (нем. *Zusammen* – вместе) либо по разные стороны (нем. *Entgegen* – напротив) по отношению к двойной связи. Стереоизомерия при асимметрическом тетраэдрическом атоме в номенклатуре ИЮПАК обозначается с использованием символов S и R (правила Кана-Ингольда-Прелога), которые означают, что обход заместителей в порядке падения приоритетности, если смот-

---

тура органических соединений: конспекты лекций, обучающие задачи и справочный материал. Казань: Изд-во Казан. гос. технол. ун-та, 2001. 336 с.

реть на атом, обладающий наименьшим приоритетом со стороны асимметрического атома, происходит против либо по часовой стрелке соответственно. Подробнее со стереоизомерией и правилами старшинства можно познакомиться в учебниках по органической химии и стереохимии.

Ни тривиальные, ни систематические имена соединений не могут эффективно использоваться для обработки компьютером, что обусловлено как сложностью алгоритмов, так и неоднозначностью самих правил кодирования. Другие достоинства и недостатки имен соединений как способа их представления изложены в табл. 3.

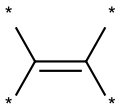
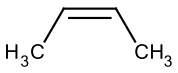
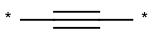
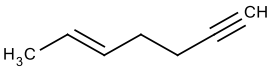
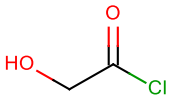
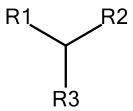
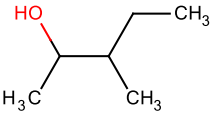
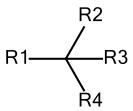
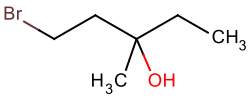
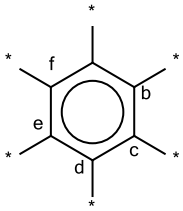
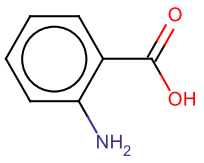
2.2.2. Линейные представления Висвессера (WLN)

Линейные представления Висвессера [20] были предложены в 1949 году [21] для того, чтобы кодировать (вручную) химические структуры при помощи коротких строк с целью их хранения на основном информационном носителе того времени – перфокартах. Эти короткие и эффективные представления быстро получили известность. На их основе уже с 1953 года были разработаны первые методы хранения и поиска химических структур [22, 23], а в начале 1960-х годов были созданы первые компьютеризированные электронные базы данных химических соединений [21, 24].

В линейных представлениях Висвессера атомы представляются при помощи символов, обозначающих соответствующие химические элементы. Кроме того, для описания некоторых часто встречающихся групп атомов и циклов используются цифровые обозначения или специальные символы (см. табл. 4). Кроме того, существуют специальные правила, которые позволяют дать уникальное имя каждому соединению (правила канонизации).

Таблица 4

Обозначения в системе WLN с примерами

Тип подструктуры	Структурный элемент	Кодирование	Пример
Водороды	H	**H**	H_2 = HH
Алкильные группы	C_nH_{2n+1}	Число ***n*** (1,2…)	CH_3- = 1, C_2H_6 = 2H
Двойная связь		**U**	 = 2U2
Тройная связь		**UU**	 = UU4U2
Амины	-NH_2, =NH, ≡N	**Z, M, N**	CH_3NH_2 = Z1; CH_3NHCH_3 = 1M1
Галогениды	F, Cl, Br, I	**F, G, E, I**	HBr = EH
Гидроксильные (тиольные) группы	-OH, -SH	**Q, SH**	CH_3CH_2OH = Q2
Простые эфиры, эфиры (тиоэфиры)	-O- -S-	**O** **S**	$CH_3CH_2OCH_3$ = 2O1 $CH_3CH_2SCH_3$ = 2S1
Карбонильная группа	>C=O	**V**	 = Q1VG
Третичный атом углерода		R1**Y**R2&R3	 = QY1&Y2&1
Четвертичный атом углерода		R1**X**R2&R3&R4	 = QX2&1&2E
Ароматические кольца		**R** – фенильное кольцо **R** B,C,D,E,F (пробел обязателен)	 = ZR BVQ

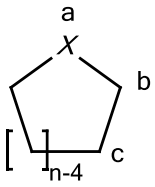
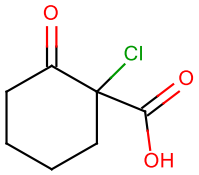
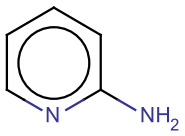
Циклы, гетероциклы		**L*n*XJ** **L** – старт цикла ***n*** – число атомов в цикле **X –** гетероатом (C не указывается) **T** – все атомы насыщенны (по умолчанию считается что все атомы цикла углерода ненасыщенные) **J –** конец описания цикла	 = L6VTJ BVQ BG  = L6NJ BZ

Таблица 5

Достоинства и недостатки линейных представлений Висвессера

Преимущества	Недостатки
✓ Легко создаются и интерпретируются человеком	○ Большое число сложных правил
✓ Короткое, емкое представление (число символов меньше числа атомов)	○ Достаточно легко сделать ошибку при кодировании
✓ Однозначные	○ Сложности автоматической конвертации в другие представления молекул
✓ С их использованием возможен текстовый поиск по подструктуре	○ При подструктурном поиске можно проводить только поиск по подструктурам, которые использованы при написании текстовой строки
✓ Включают указание стереохимии	○ Только для молекул
✓ Уникальные при следовании всем правилам	
✓ Возможно его использование для генерации фрагментов	

В дальнейшем, в связи с быстрым развитием компьютерной техники, стало возможным применять графические редакторы для ввода химических структур, а также использовать более удобные таблицы связности для обработки и хранения химической информации. В связи с этим применение WLN-представлений потеряло свою актуальность, и в настоящее время они уже практически не используются. Основным

недостатком WLN-представлений является высокая алгоритмическая сложность автоматического кодирования на компьютере химических структур и особенно правил их канонизации. Таким образом, важнейшим значением линейных представлений Висвессера является то, что они сыграли важную историческую роль в развитии хемоинформатики, явившись предшественником и в значительной мере прототипом современных типов линейных представлений молекул. В табл. 5 перечислены достоинства и недостатки WLN-представлений.

2.2.3. Линейные представления SMILES

Система линейных представлений молекул SMILES (англ. *Simplified Molecular Input Line Entry System* – система упрощенного представления молекул в строке ввода) была разработана Дэвидом Вайнингером в 1986 году в Лаборатории исследований окружающей среды США, Дулут [25, 26]. Впоследствии она получила развитие в компании Daylight Chemical Information Systems Inc., которая является владельцем товарного знака SMILES™, а также расширений системы на реакции SMIRKS™ и запросы SMART™. Программное обеспечение от компании Daylight является эталонным, хотя в настоящее время существует множество независимых реализаций SMILES в разнообразных, в том числе и в свободно распространяемых, программах. Следует, однако, иметь в виду, что зачастую программы от других производителей не поддерживают все возможности SMILES, особенно для нестандартных систем (например, для комплексов металлов, а также для нестандартных типов хиральности). SMILES является наиболее известной и широко используемой в настоящее время системой линейного представления молекул. Причиной успеха SMILES является предельная простота правил кодирования, исходя из структурной формулы. Эти правила могут быть легко запрограммированы для компьютера, при этом линейные представления SMILES легко формируются и читаются без его помощи.

В настоящее время существует множество модификаций правил SMILES для различных целей: SMARTS (для описания запросов), SMIRKS (для химических реакций), SSMILES (упрощенные SMILES), CHUCKLES (для полимеров), CHORTLES (для смесей), USMILES (уникальные, канонические SMILES), ASMILES (уникальные SMILES с указанием стереохимии), XSMILES (унифицированный формат для

обмена информацией) и др. На первых двух из них мы остановимся подробнее.

Преимущества и недостатки представления SMILES собраны в табл. 6.

Таблица 6

Преимущества и недостатки представления SMILES

Преимущества	Недостатки
✓ Создается с использованием простых и немногочисленных правил ✓ Легко создается и интерпретируется человеком и компьютером ✓ Быстрый формат обмена данными ✓ Поддерживает запросы, указание стереохимии, кодирование реакций ✓ Однозначное	○ Не уникальное (базовые правила) ○ Создание уникального представления возможно только с использованием компьютера ○ Требуется специальное указание ароматических циклов (вручную или с использованием стороннего программного обеспечения) ○ Относительно длинные – число символов немного больше числа атомов

2.2.3.1. Правила SMILES

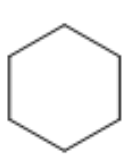
В отличие от большинства линейных представлений с достаточно сложной системой правил, система SMILES использует всего пять их основных групп [25]:

1. *Атомы*. Атомы указываются с использованием присвоенных им символов в квадратных скобках ([Au], [Se] и др.). Атомы C, O, N, S, P указываются без скобок. Атомы водорода в основном опускаются и расставляются в структурной формуле молекулы, чтобы заполнить свободные валентности. Атомы, входящие в ароматические системы, указываются прописными буквами (c, b, n, p и др.). Атомы водорода (при необходимости) и заряд атома приводятся внутри квадратных скобок ([CH], [H], [O-], [Fe2+] он же [Fe++]).

2. *Связи.* Связанные атомы в основной цепи размещаются по соседству. Одинарные связи не обозначаются, двойные обозначаются символом =, тройные – символом #. Например, этан CH_3-CH_3 обозначается как CC, этилен CH_2=CH_2 – как C=C, ацетилен CH≡CH– как C#C, синильная кислота HCN – как [H]C#N или C#N.

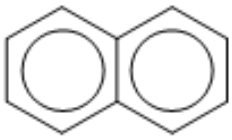
3. *Разветвления.* Ответвления от основной цепи обозначаются в круглых скобках. Например, изобутан CH_3-$CH(CH_3)$-CH_3 обозначается как CC(C)C, неопентан (2,2-диметилпропан) CH_3-$C(CH_3)_2$-CH_3 – как CC(C)(C)C.

4. *Циклы.* Циклические структуры записываются, разрывая одну или несколько из связей. Каждая разорванная связь нумеруется. Место разрыва обозначается номером связи, который следует сразу после символа атомов, при которых связь была разорвана. Только одинарная или ароматическая связи могут быть «разорваны» в цикле. Например,

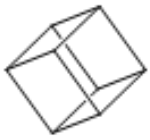


C1CCCCC1

c1ccccc1

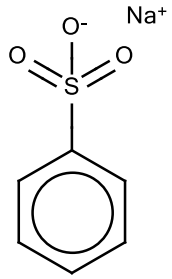


c1ccc2ccccc2c1

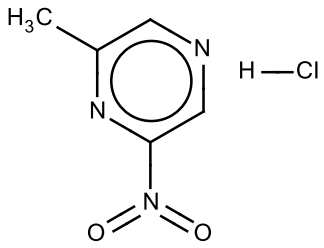


C12C3C4C1C5C4C3C25

5. *Разъединенные структуры* (т.е. состоящие из нескольких компонент, не связанных друг с другом ковалентными связями). Сюда относятся соли, межмолекулярные комплексы, не взаимодействующие соединения в смеси и другие случаи, когда нельзя говорить о ковалентной связи, соединяющей отдельные компоненты. Компоненты разъединенных структур перечисляются через точку. Например,



[Na+].[O-]S(=O)(=O)c1ccccc1
допустимо также
O=S(=O)([O-].[Na+])c1ccccc1



Cl[H].Cc1nc(cnc1)N(=O)=O
или
Cl[H].Cc1nc(N(=O)=O)cnc1

Более детальное описание системы SMILES может быть найдено на странице компании Daylight[10].

2.2.3.2. Канонические представления SMILES

Основным недостатком приведенного выше набора правил является то, что он не обеспечивает уникальность линейных представлений

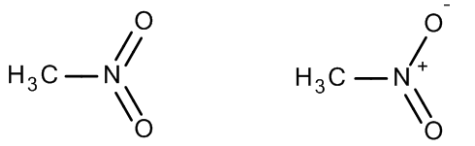
[10] www.daylight.com

SMILES, т.е. допускает формирование разных кодов для одной молекулы. Обусловлено это следующими причинами:

1. В оригинальной формулировке SMILES не существует правил приоритетности в выборе атомов, и потому линейное представление может начинаться с любого атома молекулы, что приводит к разным основным цепочкам атомов (см. Рис. 4).

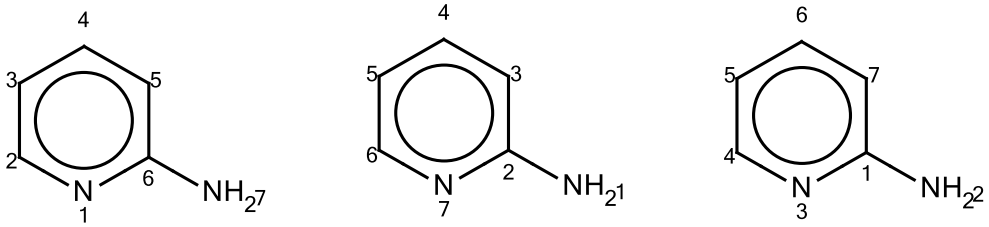
2. Определенная неоднозначность правил. При желании можно указывать или не указывать водороды или изотопы атомов, стереоизомерию атомов. Заряды, фрагменты можно указывать различным образом.

3. Системы делокализованных электронов и связей могут быть закодированы различным образом:



CN(=O)=O C[N+]([O-])=O

4. Рассмотрение разных таутомерных форм приводит к разным линейным представлениям SMILES.



n1ccccc1N Nc1ccccn1 c1(N)nccccc1

Рис. 4. SMILES, генерируемые с использованием различных нумераций атомов в молекуле 2-аминопиридина

Для преодоления этого недостатка разработаны специальные алгоритмы, которые создают уникальную нумерацию атомов, ведущую к формированию уникального кода SMILES. Каноническое наименование молекулы позволяет, преобразовав ее линейное представление в хеш-код, быстро находить в базах данных информацию о молекулах и избегать ее дублирования.

Виды канонических SMILES

Представления SMILES, сформированные на основе информации о связности молекулярного графа без учета возможности стереоизомерии, называются «обобщенными SMILES» (англ. *generic SMILES*). Для каждой молекулы существует множество вариантов обобщенного SMILES. Для того чтобы выбрать единственный канонический общий SMILES из множества возможных, используют специальный алгоритм, включающий:

☐ стандартизацию 2D-представления (т.е. вычленение стандартным образом ароматических систем внутри молекулы, выбор для нее стандартным образом таутомера и т.п.);

☐ поиск канонической нумерации вершин соответствующего молекулярного графа и создание канонического представления SMILES на его основе.

Данный алгоритм дает на выходе так называемый «уникальный SMILES» (англ. *unique SMILES*).

Общий SMILES	Уникальный SMILES
OCC	CCO
[CH3][CH2][OH]	CCO
C-C-O	CCO
C(O)C	CCO
OC(=O)C(Br)(Cl)N	NC(Cl)(Br)C(=O)O
ClC(Br)(N)C(=O)O	NC(Cl)(Br)C(=O)O
O=C(O)C(N)(Br)Cl	NC(Cl)(Br)C(=O)O

SMILES с указанием стереомерных особенностей молекулы (изомерией при двойной связи, тетраэдрическом атоме или других элементах асимметрии) и изотопного состава (если молекула обогащена теми или иными изотопами) называется «изомерным» (англ. *isomeric*). Уникальный изомерный SMILES рекомендовано называть «абсолютным SMILES» (англ. *absolute SMILES*).

Алгоритм канонической нумерации Моргана

Исторически алгоритм Моргана [27] является первым алгоритмом, который позволил проводить каноническую нумерацию атомов. Этот алгоритм первоначально был использован в химической информационной системе Американского химического общества (англ. *Chemical Abstract Service*) применительно к таблицам связанности. Он позволяет классифицировать атомы соединения с учетом существования симмет-

рично эквивалентных атомов. Например, в молекуле пропана CH_3-CH_2-CH_3 концевые метильные группы относятся к одному классу эквивалентности относительно операций симметрии в молекулярном графе[11], к другому же классу относится метиленовая группа в центре. Поэтому нет разницы в том, начинать ли название молекулы по правилам SMILES слева направо или справа налево.

Алгоритм Моргана классифицирует атомы на основе числа соседей атома (его связанности) с использованием итеративного алгоритма (релаксационный этап) с последующим применением определенных правил для создания однозначной классификации атомов в молекуле.

На первом шаге релаксационного этапа каждому неводородному атому ставится в соответствие значение его связности – число, равное количеству его соседей, не являющихся атомами водорода, т.е. степень соответствующей вершины молекулярного графа. Таким образом, у неводородных атомов метильной, спиртовой и аминной группы связность равна единице, метиленовой, амидной, эфирной и фенильной – двум и так далее. После расчета связности подсчитывается общее число классов эквивалентности k как общее число различных значений связности. На первом шаге у органических соединений каждый атом получает значение связности от 1 до 4, то есть всего атомы могут принадлежать максимум четырем классам эквивалентности (соответствующие значениям связанности 1, 2, 3 и 4), см. Рис. 5.

На втором шаге для каждого атома рассчитывается сумма значений связности всех его соседних атомов. Полученная величина носит название расширенной связности (англ. *Extended Connectivity, EC*) и, по сути, представляет собой количество связей у всех соседних атомов. Ее значение после второго шага варьируется от 2 до 12. Суммирование значений EC соседних атомов продолжается до тех пор, пока число классов эквивалентности на последующих шагах не станет постоянным или не станет уменьшаться. После достижения этого значения дальнейшие итерации не приводят к увеличению числа классов. Работа алгоритма Моргана показана на рис. 5.

---

[11] Строго говоря, под классом эквивалентности здесь понимается орбита группы автоморфизмов молекулярного графа на множестве его вершин.

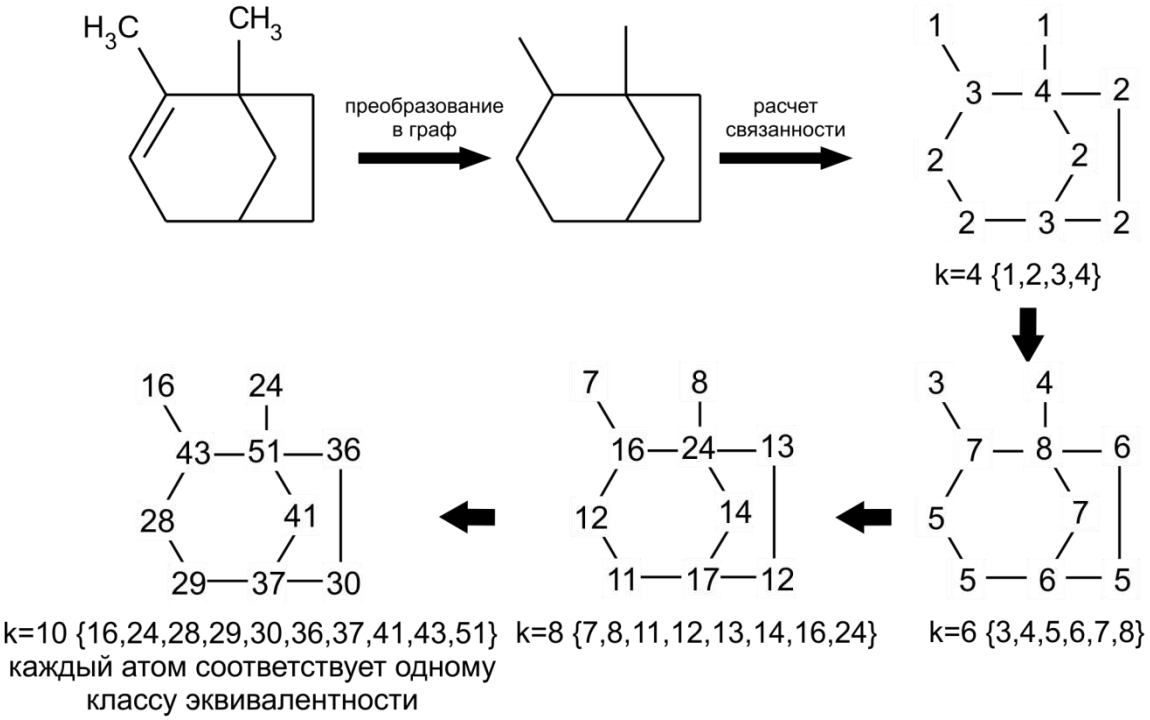


Рис. 5. Иллюстрация к использованию алгоритма Моргана. На каждом шаге указано количество классов эквивалентности. Остановка алгоритма произошла по причине равенства числа атомов числу классов эквивалентности

На следующем этапе алгоритма проводится нумерация атомов таким образом, что номер 1 присваивается атомам, имеющим максимальное значение EC, номер 2 присваивается атомам, имеющим следующее в порядке уменьшения значение EC, и так далее в порядке уменьшения ее значений. Таким образом, минимальные номера получают атомы, наиболее глубоко внедренные в структуру молекулы, максимальные номера – боковые цепи. Если два или более атомов имеют одинаковые значения расширенной связности, то их порядок приоритетности может определяться типом атома (C, N, O и т.д.), типом связи, зарядом и др. Если эти дополнительные правила не позволяют определить более приоритетный атом, то различий в представлении при любой их нумерации не будет. Основным недостатком алгоритма Моргана является его эвристический характер. В ряде случаев это приводит к ложному выявлению симметрично эквивалентных атомов, когда выявляемые классы эквивалентности не совпадают с орбитами группы автоморфизмов молекулярного графа [28, 29]. Для преодоления этой проблемы предложено несколько подходов [30, 31].

Алгоритм канонизации CANGEN

Модификация алгоритма Моргана для создания уникальных линейных представлений SMILES, разработанная в компании Daylight, называется CANGEN [26]. Она включает в себя как алгоритм канонической нумерации CANON, так и алгоритм создания уникального имени GENES.

Алгоритм CANON использует пять инвариантов графа (связность, число связей с неводородными атомами, атомный номер, знак и величина заряда, число связанных водородов). На этом основании каждому атому присваивается восьмизначное число, атомный инвариант, *AI*. На первом этапе значения *AI* сортируются, и каждому классу эквивалентности атомов присваивается ранг (порядковый номер класса, начинающийся с единицы). Ранг является аналогом параметра EC в алгоритме Моргана и пересчитывается на каждом последующем шаге итерационного процесса. После достижения согласования в значениях рангов на двух последующих шагах итерации, алгоритм GENES используется для формирования кода SMILES, начиная с атома, имеющего минимальный ранг, так, чтобы атомы следовали в строке по возможности в порядке увеличения ранга. При разветвлении направление основной цепи выбирается по атому с большим рангом. Поэтому, например, уникальный SMILES для ацетона CH_3^1-C^2O^3-C^1H_3 (верхний символ означает ранг класса) будет не CC(=O)C, а CC(C)=O.

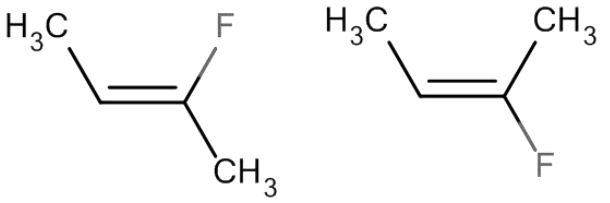
2.2.3.3. Указание стереохимии в SMILES

Стереохимическая конфигурация соединений является крайне важным их свойством, которое может существенно влиять как на биологические свойства (энантиомерия и другие виды стереоизомеров), так и на большинство физических свойств (цис-транс изомерия, диастереомерия, но не энантиомерия). Учет стереоизомерии особенно важен при работе с базами данных лекарственных или других биологически активных соединений. Поэтому зачастую указание ориентации групп при двойной связи или асимметрическом (хиральном) центре является крайне важным. Такие SMILES, как уже отмечалось, называются изомерными.

Цис-транс изомерия

Цис-транс изомерия при двойной связи указывается в линейном представлении SMILES при помощи символов / и \, которые располагаются между связанными атомами и используются как особый вид

«направленных» связей. Их применение имеет смысл, только если они указаны с обеих сторон от двойной связи. Комбинации A/X=Y/B и A\X=Y\B соответствуют транс-ориентации атомов A и B относительно двойной связи, комбинация A\X=Y/B и A/X=Y\B – их цис-ориентации (рис. 6).



C\C=C(\C)F
или
C/C=C(/C)F

C\C=C(/C)F
или
C/C=C(\C)F

Рис. 6. Примеры кодирования в SMILES ориентации относительно двойной связи

Если известна ориентация атомов не при всех двойных связях, то правила SMILES допускают частичное указание конфигурации только при двойной связи с известной ориентацией заместителей.

Стереоизомерия при тетраэдрическом атоме

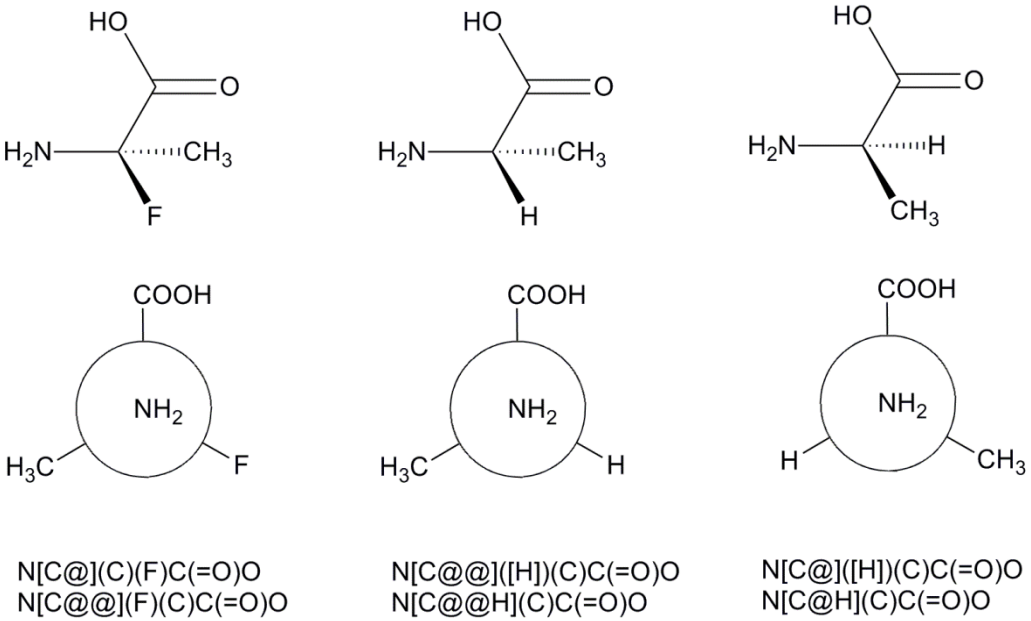
Расположение четырех атомов при тетраэдрическом центре указывается с использованием символа @ (против часовой стрелки, что совпадает с написанием буквы) или @@ (по часовой стрелке). Символ указывается после хирального атома и заключается вместе с атомом в квадратные скобки.

Направление по или против часовой стрелки определяется, исходя из следующих соображений:

☐ Молекула располагается так, чтобы наблюдатель смотрел от атома, стоящего в строке SMILES перед асимметрическим атомом, к асимметрическому атому (рис. 7);

☐ Если направление обхода трех атомов, указанных в SMILES сразу после асимметрического атома, в порядке их следования в строке происходит против часовой стрелки, то асимметрическому атому присваивается символ @, если по часовой стрелке – то @@.

При наличии атома водорода при асимметрическом атоме он указывается в явном виде в квадратных скобках первым после асимметрического центра (рис. 7).



N[C@](C)(F)C(=O)O
N[C@@](F)(C)C(=O)O

N[C@@]([H])(C)C(=O)O
N[C@@H](C)C(=O)O

N[C@]([H])(C)C(=O)O
N[C@H](C)C(=O)O

Рис. 7. Определение написания изомерного SMILES при наличии хирального тетраэдрического атома

Другие виды стереоизомерии также кодируются в SMILES. Их описание можно найти на официальном сайте Daylight по адресу: http://daylight.com/dayhtml/doc/theory/theory.smiles.html

2.2.3.4. Линейное представление реакций (SMIRKS)

SMIRKS является расширением SMILES для описания реакций. С этой целью приводится *отображение атомов* (англ. *atom mapping*), указывающее, какие атомы продуктов реакции соответствуют каким атомам реагентов. Для этого атомы реагентов и продуктов должны быть пронумерованы, и соответствие между ними должно быть указано с обеих сторон от стрелки реакции (рис. 8). Таким образом, при задании реакций надо представлять, как происходит реакция, какие атомы и как изменяются. При этом часть атомов меняют свое ближайшее окружение за счет разрушения старых и образования новых связей, другая часть атомов остается в том же окружении.

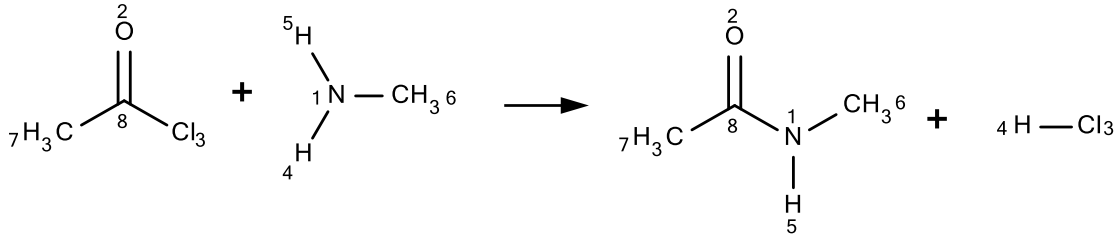


Рис. 8. Отображение атомов на примере реакции аминирования ацетилхлорида

Для описания реакций система SMILES была расширена следующими правилами:

1. Реагенты и продукты реакций должны содержать одинаковое число и одинаковый тип отображаемых атомов, причем каждому отображаемому атому в реагентах должен соответствовать строго один атом в продуктах реакции, а каждому отображаемому атому в продуктах реакции – один атом в реагентах, т.е. отображение атомов должно быть попарным и взаимно однозначным. Отображения атомов указывают с помощью чисел после символа атома через двоеточия, помещая отображаемые атомы в квадратную скобку, например, [C:1], [H:100]. Номера не обязательно должны быть последовательными.

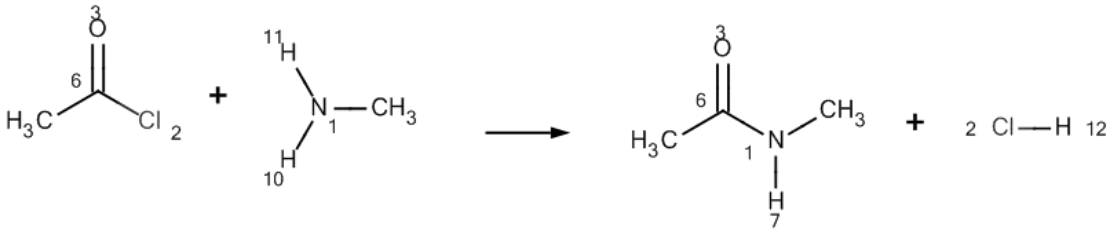
2. Стехиометрия атомов и молекул в реакции полагается равной 1:1. При неэквимолярной стехиометрии реакции необходимо указать соответствующие молекулы несколько раз явным образом.

3. Атомы водорода, которые изменяют свое окружение в результате реакции, должны быть указаны явно. Указанные атомы водорода в продуктах реакции должны быть отображены на соответствующие атомы реагентов.

4. Атомы и связи в продуктах и реагентах указываются так же, как и в обычных SMILES.

5. Символом реакции (трансформации) является знак >. Необходимо соблюдать следующий синтаксис при представлении реакции: реагент1. реагент2 > низкомолекулярный реагент1. низкомолекулярный реагент2 > продукт1. продукт2. Низкомолекулярные реагенты (например, [H+] – кислоты, OS(=O)(=O)O – серная кислота, O – вода, [H][H] – водород) могут быть перечислены среди основных реагентов или продуктов реакции, либо пропущены. Тогда реакция записывается следующим образом: реагент1. реагент2 >> продукт1. продукт2.

Пример линейного представления реакции приведен на рис. 9.



C[C:6]([Cl:2])=[O:3].[H:11][N:1]([H:10])C>>[H:7][N:1](C)[C:6](C)=[O:3].[Cl:2][H:12]

Рис. 9. Отображение атомов и соответствующий SMIRKS для реакции аминирования ацетилхлорида

2.2.3.5. Представление шаблонов для спецификации молекулярных фрагментов (SMART)

SMARTS (от англ. *SMiles Arbitrary Target Specification*) – это основанный на SMILES язык, позволяющий специфицировать фрагменты (подструктуры) внутри молекул при помощи шаблонов. Он широко используется в хемоинформатике для следующих целей:

- формирования запросов для подструктурного поиска в химических базах данных;
- описания аддитивных схем расчета разнообразных свойств химических соединений;
- спецификации фильтров с целью отфильтровывания высокореакционных, токсичных и других нежелательных соединений при виртуальном скрининге химических баз данных (см. [32, 33]);
- задания правил стандартизации химических структур (путем спецификации стандартных ионизационных, таутомерных и мезомерных форм) при подготовке химических баз данных к виртуальному скринингу, а также выборок для SAR/QSAR/QSPR-анализа (см. [34]);
- задания правил сегментации химических структур, например, с целью выявления т.н. «привилегированных фрагментов» (см. [35, 36]);
- спецификации фармакофорных центров (см., например, [37]);
- спецификации типов атомов при задании разнообразных параметризаций силовых полей и т.п.

Правила SMART позволяют задавать практически любой фрагмент молекулы, используя правила SMILES. Кроме того, они дают возможность создавать сложные логические запросы, содержащие условия (например, атом A не должен быть азотом; атом A должен быть или азотом или углеродом, и т.п.). Логические операции задаются следующим образом:

- если условия (свойства) перечислены через запятую, то достаточно, чтобы выполнялось одно из них (A,B,C = A или B или C), т.е. запятая означает логическую связку ИЛИ;

- если условия перечислены через символ &, то все они должны выполняться (A&B = A и B), т.е. знак & означает логическую связку И;
- наличие символа ! перед каким-либо условием (свойством) означает необходимость того, чтобы оно не соблюдалось (!A = не A), то есть символ ! означает логическую операцию отрицания.

Шаблон, сформированный по правилам SMART, может содержать спецификацию как атомных, так и связевых примитивов. Атомные примитивы обычно помещаются в квадратные скобки (если не специфицируют просто подструктуру). Некоторые из них перечислены в табл. 7.

Таблица 7

Атомные примитивы в языке SMARTS

Символ	Описание символа	Умолчание
*	Любой атом	нет
A	Любой ароматический атом	нет
A	Любой алифатический атом	нет
D<n>	Степень узла: <n> указанных явно связанностей при данном узле (неявные атомы водорода не учитываются)	1
H<n>	Число атомов водорода: <n> указанных явно (в SMILES) атомов водорода, связанных с атомом	1
h<n>	Число атомов водорода: <n> не указанных явно (в SMILES) атомов водорода, связанных с атомом	как минимум 1
r<n>	Участие в цикле размера <n>	атом в любом цикле
v<n>	Валентность: суммарный порядок связей <n>	1
X<n>	Связанность: суммарное число соседей <n> (без учета порядка связи)	1
- <n>	Отрицательный формальный заряд величиной <n>	-1
+<n>	Положительный формальный заряд величиной <n>	+1
#n	Атомный номер<n>	нет

@	Ориентация заместителей против часовой стрелке при тетраэдрическом атоме
<n>	Атомная масса элемента

SMARTS = [OH]c1ccccc1&C[OH]

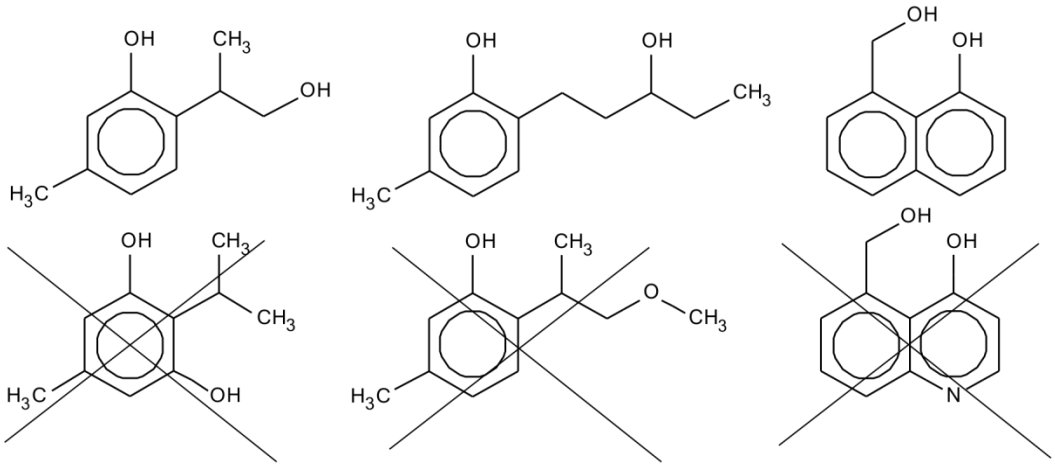


Рис. 10. Поиск в базе данных из шести соединений по приведенному шаблону SMARTS даст совпадения только с тремя

Например, [OH]c1ccccc1&C[OH] – запрос на поиск молекулы, содержащей фрагменты фенола и алифатического спирта (рис. 10), [!v4] – найти атомы, валентность которых отличается от 4, [R3, R4] – найти атомы, входящие в трехчленные или четырехчленные циклы, N[C@@H](C)C(=O)O – найти фрагмент S-аланина.

Связевые примитивы указывают на свойства связей в молекулах. Одинарная связь обозначается как «-», а отсутствие обозначений связи означает одинарную или ароматическую связь. Остальные типы связей специфицируются в SMART так же, как и в SMILES (= двойная, # тройная). Некоторые связевые примитивы перечислены в табл. 8.

Таблица 8

Связевые примитивы в языке SMARTS

Символ	Описание символа
/	Связь, направленная «вверх»
\?	Связь, направленная «вниз» или не обозначенная
:	Ароматическая связь
~	Любая связь
@	Любая связь в цикле

Например, шаблон c-c указывает на ароматические атомы углерода, связанные одинарной связью (например, между кольцами бифенила C_6H_5-C_6H_5), шаблон [N+]~*~*~[O-] позволяет найти молекулы, в которых положительно заряженный атом азота и отрицательно заряженный атом кислорода разделены пятью связями (например, в цвиттерионе глицина NH_3^+-CH_2-CO-O^-).

Другие правила SMART приведены на сайте компании Daylight: http://daylight.com/dayhtml/doc/theory/theory.smarts.html.

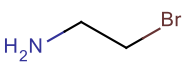
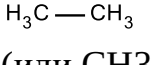
2.2.4. Линейные представления SLN

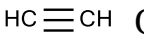
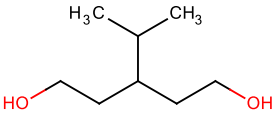
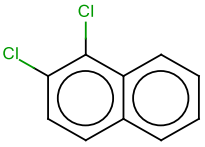
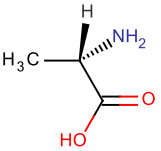
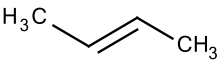
Линейное представление SLN (от англ. *Sybyl Line Notation*) позволяет кодировать структуры органических и неорганических молекул (а также их анионов, катионов и радикалов), полимеров, макромолекул и комбинаторных библиотек, а также запросы для поиска в химических базах данных, структуры Маркуша, и химические реакции [38, 39].

SLN было разработано компанией Tripos Inc. с целью решить ряд проблем SMILES, связанных, в частности, с трактовкой ароматичности как атомного свойства, а также неявным учетом атомов водорода. SLN является одним из основных форматов, используемых в программном обеспечении компании Tripos (UNITY, CONCORD, SYBYL, Galahad, Tuplets, Benchware Data Miner). Основным отличием SLN от SMILES является явное указание атомов водорода в молекуле и введение специального символа для ароматических связей. Благодаря этому представление SLN выглядит весьма похожей на конденсированную структурную формулу. Кратко правила SLN представлены в табл. 9.

Таблица 9

Обозначения в системе SLN

Структурный элемент	Обозначение	Пример
Атомы	Обозначаются символами элементов. Число атомов обозначается цифрами. Атомы водорода обозначаются после связанного с ними атома	CH_4 = CH4  = NH2CH2CH2Br
Связи	Одинарная связь – символ «–» или не обозначается. Двойная связь – символ «=». Тройная связь – сим-	 = CH3CH3 (или CH3-CH3) = CH#CH

	вол «#». Ароматическая связь – символ «:». Несвязанные атомы или связи с неизвестным порядком – символ «.»	 CH3ONa = CH3O.Na
Ветвления	Ответвления от основной цепи берутся в круглые скобки	 OHCH2CH2CH(CH(CH3)CH3)CH2CH2OH
Циклы	Начало цикла – [*n*] Конец цикла – @*n* *n* – любое число (не обязательно по порядку)	 ClC[1]:C(Cl):CH:CH:C[2]:CH:CH:CH:CH:C:@1:@2
Стереохимия	Обозначается атрибутами: При тетраэдрическом центре – атрибут атома [S=...], возможные значения: U (неизвестно), S или R (по Кану-Ингольду-Прелогу), N или I (если поворот от первого к третьему атому после указанного центра происходит по часовой стрелке – N, иначе I), D или L (D-L номенклатура – по отношению к глицериновому альдегиду). При двойной связи – атрибут связи [S=...], возможные значения C или T (цис-транс), E или Z (E,Z-номенклатура по Кану-Ингольду-Прелогу) , N или I (если два атома указанные первыми перед и после двойной связи, расположены в транс-позициях – N, иначе I).	 CH3C[S=I]H(NH2)C(=O)OH  CH3CH=[s=i]CHCH3 или CH3CH=[s=n]C(CH3)H

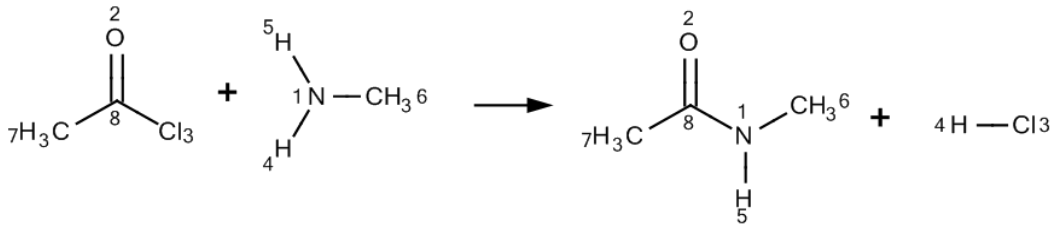
SLN также позволяет указывать специфические свойства атомов, такие как заряд, изотопный состав, стереохимическая конфигурация, а также связи с использованием специальных атрибутов. Они заключаются в квадратные или треугольные скобки и размещаются непосред-

ственно после обозначений атомов или связей, к которым относятся. Наиболее важные среди них:

- Стереохимическая конфигурация при атоме указывается с использованием атрибута [s=], возможными вариантами значений которого являются S или R (правила Кана-Ингольда-Прелога), D или L (тривиальная номенклатура), N или I (внутренняя номенклатура SLN), а также U (расположение атомов неизвестно). Например, S-форма аланина: NH2C[s=S]H(CH3)C(=O)OH или NH2C[s=N]H(CH3)C(=O)OH;
- Стереохимическая конфигурация при двойной связи обозначается также атрибутом [s=], однако в данном случае его значениями могут быть E или Z (правила Кана-Ингольда-Прелога), N или I (внутренняя номенклатура SLN), а также U (расположение атомов неизвестно). Например, для транс-бутена допустимы следующие названия: CH3CH=[s=E]CHCH3, CH3CH=[s=I]CHCH3, CH3CH=[s=N]C(CH3)H;
- Заряд атома указывается как [+n] или [–n] (n – величина заряда), или с использованием атрибута [charge=]. Пример: CH3C(=O)O[–].Na[+].

SLN позволяет также указывать «макроатомы», например *Monomer_1, Ala, His*, для краткой записи структуры пептидов, полимеров и макромолекул. Например, трипептид аланил-глицил-гистидин можно записать таким образом: HAlaGlyHisOH.

Способы задания шаблонов для запросов в SLN и SMART похожи, но с некоторыми синтаксическими особенностями: «любой атом» задается в SLN словом Any, в шаблонах могут присутствовать атрибуты, причем тип связи обозначается атрибутом [type=]. Кроме того, отличается синтаксис логических операций.



CH3[#7]C[#8](=O[#2])Cl[#3]+N[#1]H[#5]H[#4]C[#6]H3->\
\CH3[#7]C[#8](=O[#2])N[#1]H[#5]C[#6]H3+Cl[#3]H[#4]

Рис. 11. Пример кодирования химической реакции при помощи линейного представления SLN

Кодирование реакций требует так же, как и в SMIRKS, отображения атомов, изменяющих свое окружение с использованием атрибута [#n], а также указания связей изменяющих свой порядок с использованием атрибута [rc=]. Последний атрибут может принимать значения: n – связь не изменяется (указывать не обязательно); c – связь изменяется при реакции; x – связь образуется или разрывается (см. рис. 11).

Детально ознакомиться с современными правилами SLN можно по публикациям [38, 39].

По сравнению со SMILES представление SLN выглядит более сложным, и строка у SLN обычно длиннее. С другой стороны, более детальное описание молекул при помощи SLN позволяет избежать неоднозначности. В целом, если не обращать внимание на более длинную строку, SLN более совершенен, универсален и удобен, чем SMILES. Достоинства и недостатки SLN приведены в табл. 10.

Таблица 10

Достоинства и недостатки SLN

Преимущества	Недостатки
✓ Простой способ кодирования	○ Не уникальное (базовые правила)
✓ Небольшое количество простых правил	○ Создание уникального представления возможно только с использованием компьютера
✓ Легко создается и интерпретируется человеком и компьютером	○ Требуется специальное указание ароматических циклов (вручную или с использованием стороннего программного обеспечения)
✓ Похожи на рациональные формулы в полуразвернутом виде	○ Относительно длинные – число
✓ Быстрый формат обмена данными	
✓ Поддерживает обобщенные	

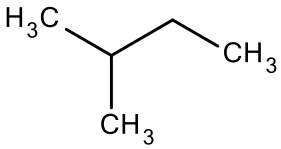
структуры, запросы, указание стереохимии, кодирование реакций, макроатомы (атомные группировки и молекулы) ✓ Однозначное	символов немного больше числа атомов

2.2.5. Идентификатор InChI

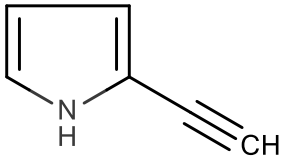
Международный химический идентификатор ИЮПАК InChI (англ. *INternational CHemical Identifier*) является некоммерческим продуктом, разработанным в 2000-2004 годах в рамках специального проекта ИЮПАК с привлечением специалистов из различных компаний. Целью данного проекта было создание бесплатного, универсального и удобного языка кодирования структур химических веществ для баз данных и обмена информацией. Помимо этого, разработан специальный алгоритм уникального преобразования InChI в хэш-коды фиксированной длины (27 символов), называемые InChIKey.

Развитие InChI в настоящее время проводится некоммерческим сообществом InChI Trust под руководством подкомитета ИЮПАК по InChI. Периодически выпускается бесплатное программное обеспечение для конвертации молекул в линейное представление InChI, InChIKey и наоборот, причем доступны также и исходные коды программ. Все это способствует широкому распространению InChI. Благодаря этому многие программы поддерживают данный формат записи структуры молекул. Более того, представление InChI может использоваться для поиска соединений с помощью поисковой системы Google.

Общедоступные web-интерпретаторы InChI, поддерживающие множество форматов ввода-вывода, доступны на базе портала ChemSpider, http://www.chemspider.com/inchi-resolver/, а также на сайте Национального ракового института США, http://cactus.nci.nih.gov-/chemical/structure. Детальное описание языка, программное обеспечение и исходные его коды для интерпретирования InChI можно найти на сайте ИЮПАК http://www.iupac.org/home/publications/e-resources-/inchi.html и общества InChI Trust www.inchi-trust.org.



InChI=1S/C5H12/c1-4-5(2)3/h5H,4H2,1-3H3
InChIKey=QWTDNUCVQCZILF-UHFFFAOYSA-N



InChI=1S/C6H5N/c1-2-6-4-3-5-7-6/h1,3-5,7H
InChIKey=IBHAUHPHCOMUJR-UHFFFAOYSA-N

Рис. 12. Примеры InChI и InChIKey

Строка InChI всегда начинается с фразы «InChI=». Фраза «InChIKey=» в начале строки означает, что приводится 27-значный хэш-код InChI, разделенный двумя знаками тире на 3 блока (рис. 12). Сразу после фразы «InChI=» приводится номер версии InChI, использованной для его создания[12], причем символ S означает стандартный InChI, созданный специальной программой. Далее, после знака / следует формула соединения в соответствии с правилами Хилла (см. выше). InChI имеет несколько «слоев», разделенных символом «/». Идущий после него символ обозначает вид слоя: /c (связанности атомов), /h (количество атомов водорода). Эти слои являются обязательными. При необходимости приводятся также другие слои: /q (заряд), /b (стереохимия двойных связей), /t (стереохимия при асимметрическом тетраэдрическом атоме) и некоторые другие. Далее следует описание соответствующих характеристик молекул. Нумерация атомов, кроме атомов водорода, идет в порядке встречаемости их в формуле Хилла (например, в молекуле спирта C2H6O – атомы углерода будут иметь номера 1 и 2, O – номер 3). Номера одинаковых атомов (например, углерода) присваиваются атомам программами в соответствии с внутренними (открыто не документированными) правилами.

В слое связанности указывается, какие атомы связаны друг с другом с использованием символа «–» при обходе всех атомов молекулы по длиннейшему пути. Ветвления обозначаются скобками, причем при

[12] В настоящее время существует только первая версия.

наличии циклов номера атомов могут повторяться несколько раз. Номера атомов, с которыми связаны атомы водорода (*n*) и число атомов водорода (*m*), связанных с каждым атомом, обозначаются в слое, начинающемся с /h в виде *n*H*m* (рис. 13).

Отметим, что информация о типе химических связей в InChI не кодируется. Это дает возможность кодировать в стандартном InChI все возможные таутомеры молекулы (рис. 13). Для этого в слое /h подвижные атомы водорода, которые могут принадлежать разным атомам в различных таутомерных формах, записываются строкой вида (Hn, 1, 2…m), где n – число подвижных атомов водорода, 1, 2…m – номера тяжелых атомов, с которыми они могут быть связаны в различных таутомерах. Отдельные таутомерные формы также можно представлять с помощью InChI (рис. 13). Для этого расположение атомов водорода, характерное для данной формы, записывается в специальном слое зафиксированных атомов водорода, начинающегося маской /f/h (полученный InChI уже не будет стандартным).

Достоинства и недостатки представления InChI приведены в табл. 11.

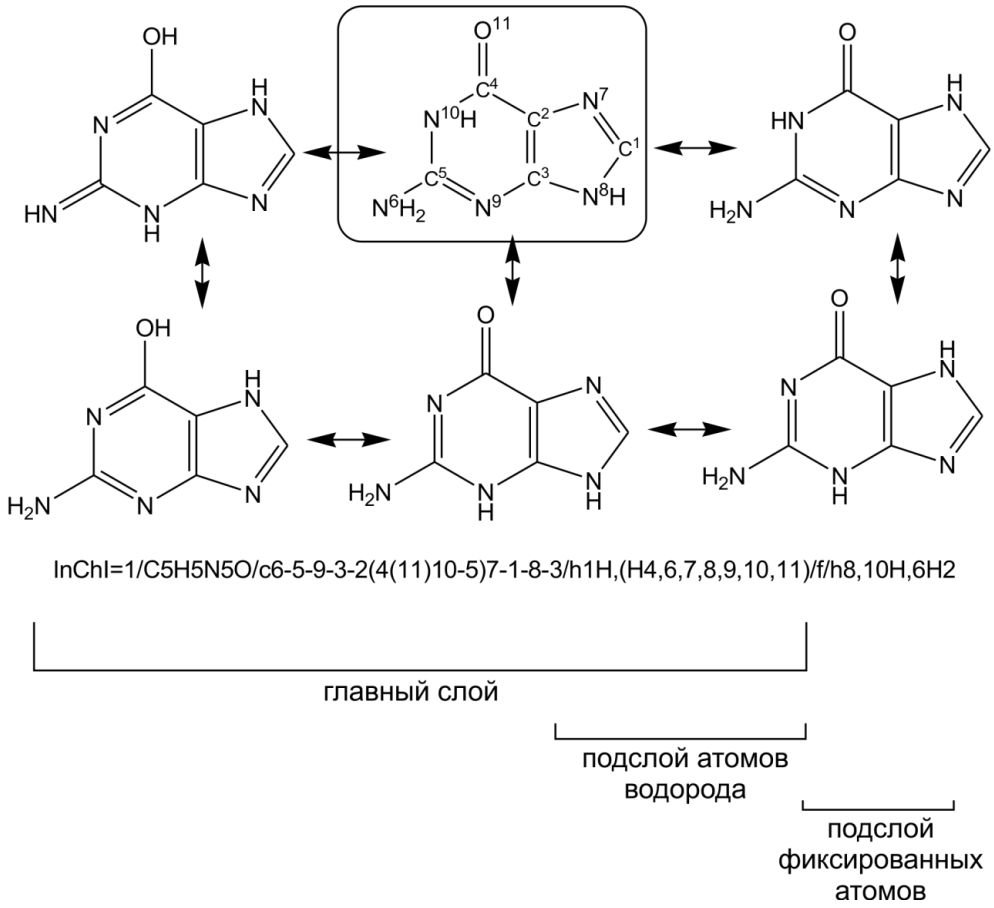


Рис. 13. Кодирование одного из таутомеров (обведен) молекулы гуанина с помощью InChI. Если оставить только главный слой, то полученный стан-

дартный InChI будет общим для всех таутомерных форм молекулы, некоторые из которых приведены. Надстрочные символы в обведенном таутомере обозначают стандартную нумерацию атомов в InChI для данной молекулы.

Таблица 11

Достоинства и недостатки InChI

Достоинства	Недостатки
✓ Уникальность ✓ Однозначность ✓ Нет необходимости вводить правила для выявления ароматических атомов или связей ✓ Учет стереоизомерии ✓ Все таутомерные и резонансные формы молекулы кодируются в одном представлении ✓ Уникальное программное обеспечение с открытым кодом	○ Не наглядны ○ Сложно создавать вручную ○ Только для молекул ○ Правила нумерации атомов в документации не определенны

2.3. ПРЕДСТАВЛЕНИЯ ГРАФОВ

С математической точки зрения, структурная формула химического соединения является графом. Графы представляют молекулу в виде набора *вершин*, соответствующих атомам, и *ребер*, их соединяющих, которые соответствуют химическим связям[13]. Графы показывают, как атомы связаны друг с другом или, иначе говоря, отражают *топологию* молекулы. Рассмотрим подробнее, что под этим подразумевается.

Все молекулы в той или иной степени являются гибкими за счет постоянно происходящего внутреннего движения[14], которое заключается в колебаниях вдоль ковалентных связей, «ножничных» движений связанных троек атомов, приводящих к изменениям валентного угла между ними, а также «внутреннего вращения» вокруг ординарных связей, приводящего к изменению соответствующих торсионных углов. Характерной особенностью внутреннего движения молекул является

[13] С математической точки зрения, граф – это совокупность непустого множества вершин и множества пар вершин (ребер).

[14] Внутреннее движение в определенной мере происходит даже при температуре абсолютного нуля.

то, что деформация их пространственной формы происходит непрерывно и не сопровождается разрывами ковалентных связей между атомами. Раздел математики, изучающий свойства объектов, которые остаются неизменными при непрерывных деформациях, называется топологией. Одним из таких свойств является связность молекулы, которая как раз и описывается графом. Следовательно, граф однозначно идентифицирует топологическое свойство молекулы – связность. На несколько менее строгом языке, принятом у химиков, это звучит так: молекулярный граф описывает (отражает) топологию молекулы.

Простой граф не содержит информации о типе вершин и ребер – он отражает только связность. В случае простейших алканов, когда отсутствуют гетероатомы и кратные связи, а атомы водорода могут быть опущены (поскольку число присоединенных атомов водорода к атому углерода может быть найдено по правилу валентности), простые графы достаточно полно описывают структуру молекул. В более общем случае, когда молекула содержит гетероатомы (т.е. атомы, не являющиеся углеродом и водородом) и кратные связи, для описания молекул используют *помеченые графы*, в которых вершины несут *метки*, соответствующие типу химического элемента соответствующего атома, а ребра – метки, кодирующие формальный порядок связи соответствующего ребра. При необходимости, вершины молекулярного графа могут также нести дополнительные метки для указания формального заряда на атоме, типа его изотопа, наличия свободного радикала и др., а дополнительные типы меток на ребрах могут быть использованы для кодирования химических реакций, структур полимеров и др.

Таким образом, в большинстве случаев молекулярный граф с мечеными вершинами и ребрами достаточно адекватно представляет структуру молекулы. Теория графов весьма хорошо разработана в математике. Там не менее у представления молекул при помощи только одних графов есть существенный недостаток – с их помощью не может быть адекватно описана стереоизомерия. Для случаев стереоизомерных молекул изображение графа на плоскости (т.н. *геометрический граф*) обычно дополняют *топографической* информацией (например, в виде изображений клинышек), однако и этого не всегда бывает достаточно для точного представления стереоизомера. Таким образом, информации, содержащейся в графе, недостаточно для описания стереохимии химического соединения. Причина этого заключается в том, что, как было показано в ряде работ в области математической химии,

адекватное описание стереоизомерии может быть достигнуто только путем явного рассмотрения отношений, охватывающих четверки атомов. Действительно, правила присваивания обозначений Z и E конфигурации при двойной связи, а также S и R конфигурации хирального центра по правилу Кана-Ингольда-Прелога, основаны на рассмотрении четверок атомов. Следовательно, для адекватного описания стереохимических особенностей молекул представление вершинно- и реберно-помеченного молекулярного графа должно быть дополнено информацией, относящейся к упорядоченным наборам из четверок атомов.

Следует отметить, что существует математический аппарат, позволяющий с единых позиций адекватно описывать и перечислять структуры молекул, их стереоизомеры, а также органические реакции. Им является основанный на комбинаторной теории групп подход, называемый «*лестница комбинаторных объектов»,* и разработанный С.С. Трачом и Н.С. Зефировым [40, 41].

Существуют различные компьютерные представления молекулярных графов. Так, связность вершин графа и ряд других его характеристик могут быть описаны с помощью специальных типов матриц или таблиц. Граф можно также представить путем перечисления входящих в него подграфов. Полученные представления могут иметь вид вектора (одномерного массива данных), квадратной или прямоугольной матрицы.

Первые представления графов в хемоинформатике начали использовать примерно с начала 1960-х годов. Их разработка была связана с необходимостью эффективной автоматической кодировки структуры молекулы в машинно-читаемый код (что было невозможно с существующим в то время представлением WLN), и обеспечения возможности вести поиск не только по структуре, но и по подструктуре.

2.3.1. Векторное представление

Векторное представление описывает молекулу в виде вектора (одномерного массива) чисел. Эти числа могут быть целыми (например, 3), действительными (например, 3.14), или единицами и нулями (битами). Последний вид представления называется битовым.

2.3.1.1. Битовое представление молекулы

Битовая строка (также называемая битовым вектором, битовой картой (англ. *bitmap*) или булевым массивом) представляет собой

строку, состоящую из цифр 0 или 1, например, 0100011. Последние являются разрядами числа, представленного в двоичной системе[15] исчисления, и вместе называются битами (от англ. *Binary digiT* – двоичное число). Бит может рассматриваться как минимальная единица информации, свидетельствующая о том, что какая-то характеристика присутствует (1) или отсутствует (0). По этой причине битовые вектора могут использоваться для кодирования молекулярной структуры (молекулярного графа) как совокупности фрагментов, ее составляющих; т.е. если фрагмент присутствует в структуре – то в соответствующем элементе массива мы ставим единицу, если нет – нуль.

Битовое представление молекулы позволяет проводить операции крайне быстро. Так как архитектура компьютера оптимизирована под двоичную систему, любое число, слово, рисунок – все в компьютере в конечном итоге превращается в двоичный массив, с которым уже и происходят операции. Именно использование таких представлений позволяет быстро проводить поиск молекул даже в таких крупнейших базах данных, как SciFinder или Reaxis.

Существует три вида битовых представлений молекулы: структурные ключи, молекулярные отпечатки и хешированные молекулярные отпечатки.

Созданные первоначально для ускорения поисков по подструктуре и молекулярному сходству, битовые представления молекул стали особенно популярными для использования в качестве дескрипторов в интеллектуальном анализе данных, особенно для анализа «сходства» молекул по структуре и для кластеризации.

Самым существенным недостатком битового представления является то, что непосредственно из него практически невозможно восстановить структуру молекул.

---

[15] В двоичной системе исчисления существует только два числа – 0 или 1 (в десятеричной, которую обычно мы используем, их 10: 0,1,2,…9), поэтому числа, начинающиеся с 3, записываются в несколько разрядов. Например, число 0 в десятеричной системе исчисления представляется в двоичной системе как 0, 1 – как 1, 2 – как 10, 3 – 11, 4 – 100 и т.д. Таким образом, каждый разряд в двоичной системе исчисления содержит или 0, или 1, а любое число есть набор нулей и единиц.

Структурные ключи

Структурные ключи широко используются для ускорения поиска по подструктуре начиная с 1971 года [42]. **Структурный ключ (англ. *structure key*) – это одномерный булевый массив (битовая строка), в котором каждый из элементов (разряд двоичного числа) означает присутствие (ИСТИНА или 1) или отсутствие (ЛОЖЬ или 0) заранее определенного фрагмента** (рис. 14).

Такими фрагментами могут быть:

- какой-либо химический элемент в определенном количестве, например, «присутствие по меньшей мере 4 атомов азота»;
- определенные циклы, например, бензольный, циклогексановый, циклопропановый;
- те или иные функциональные группы, например, гидроксильная, карбоксильная, аминная;
- те или иные группировки атомов (пары, цепочки, атомы в циркулярном окружении других атомов), например: C-N; углерод, окруженный тремя атомами кислорода; любой атом в окружении трех азотов;
- редкие фрагменты (изотопы, атомы со стереохимическими особенностями, атомы с валентностью больше 5 и др.).

Важным свойством структурных ключей является требование наличия библиотеки базовых фрагментов, которые определяют, какому фрагменту какой бит соответствует в строке (рис. 14). Хотя имеется принципиальная возможность использовать наличие любого фрагмента химической структуры в качестве ключа, некоторые из наборов, доказавших на практике свою эффективность, стали стандартными в хемоинформатике: ключи MACCS [43] компании MDL, длиной 166 или 960 бит, ключи BCI, скрининговые векторы CACTVS [44]. Необходимо, однако, помнить, что реализация структурных ключей в разных программах может отличаться. Это связано с тем, что библиотеки фрагментов для этих структурных ключей описаны в литературе не полностью или неоднозначно, либо вообще не опубликованы в открытой печати, поскольку являются разработкой частных компаний.

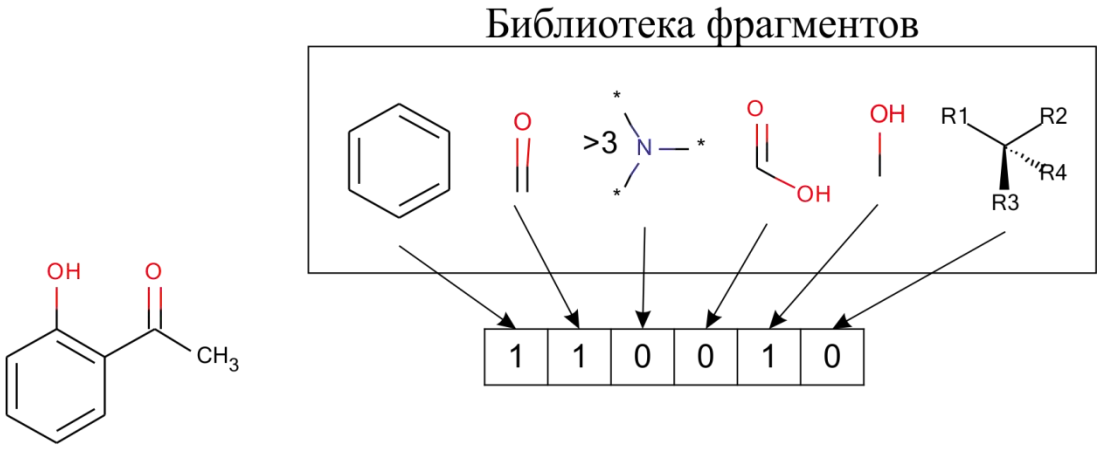


Рис. 14. Создание структурного ключа

Операции со структурными ключами позволяют проводить быстрый поиск по структуре, подструктуре и подобию, однако само создание битовой строки является трудоемкой задачей, так как требует поиска всех подструктур из библиотеки фрагментов (обычно несколько сотен фрагментов) для всех молекул в базе данных.

Правильный выбор фрагментов оказывает существенное влияние на эффективность работы с данным представлением. Неоптимальный выбор библиотеки фрагментов структурных ключей может привести к тому, что базы данных будут ими плохо описываться. Причиной этого является то, что родственные молекулы могут иметь много общих фрагментов (много единиц в одних и тех же положениях битовой строки), не содержать одни и те же фрагменты (много нулей в одних и тех же положениях), но при всем этом в библиотеке отсутствуют фрагменты, позволяющие различать эти молекулы (если положения в битовой строке для них не предусмотрено). В результате этого, с одной стороны, ключ «не работает» полностью, так как остается много фрагментов с нулевой встречаемостью (много нулей в строке), а с другой стороны, различить молекулы на основании ключей невозможно (то есть возникают проблемы с *дискриминацией молекул*). Чтобы избежать этого, необходимо вводить большое число разнообразных фрагментов и увеличивать размер структурного ключа, из-за чего его эффективность еще больше падает (процент нулей в строке к общей длине строки еще больше уменьшается) и увеличивается время создания структурного ключа.

Наиболее эффективными будут структурные ключи со схожей частотой встречаемости единиц и нулей в строке в разных положениях, и при этом с одинаковой встречаемостью каждого конкретного фрагмен-

та из библиотеки во всех соединениях. Это достигается за счет определения и замены фрагментов, которые встречаются слишком часто или слишком редко – то есть оптимизации структурного ключа. Такой способ используется, например, в CAS Registry. Это ограничивает возможности применения библиотек «стандартных» структурных ключей для конкретных нужд.

Создание универсального структурного ключа является задачей крайне сложной и, по-видимому, неразрешимой, поскольку эффективность структурного ключа в значительной мере зависит от используемой библиотеки фрагментов, типичных запросов к базе данных, а также классов соединений, входящих в ее состав. Пополнение базы данных соединениями, содержащими новые фрагменты, потребует пересмотра библиотеки фрагментов структурного ключа.

Молекулярные отпечатки

Молекулярные отпечатки (т.н. «фингерпринты», от англ. *fingerprints*), так же как и структурные ключи, описывают наборы присутствующих в молекулах структурных фрагментов. Для создания молекулярных отпечатков библиотека фрагментов, однако, определяется не заранее, а генерируется для данной конкретной базы данных.

На первом этапе соединения в базе данных «разбиваются» на все возможные структурные фрагменты, которые могут быть:

- атомами;
- цепочками атомов и связей (*sequences*) длиной от 2 до некоторого числа (часто не больше 7);
- атомами со своим ближайшим окружением (*augmented atoms*).

Для молекулы со SMILES OC(C)C#N, например, такими фрагментами будут[16]:

- атомы: O, C, N
- цепочки: OC, CC, C#N, OCC, CC#N, CCC, OCC#N, CCC#N
- окружения: C(C)(O)C.

[16] Приведены только не совпадающие значения

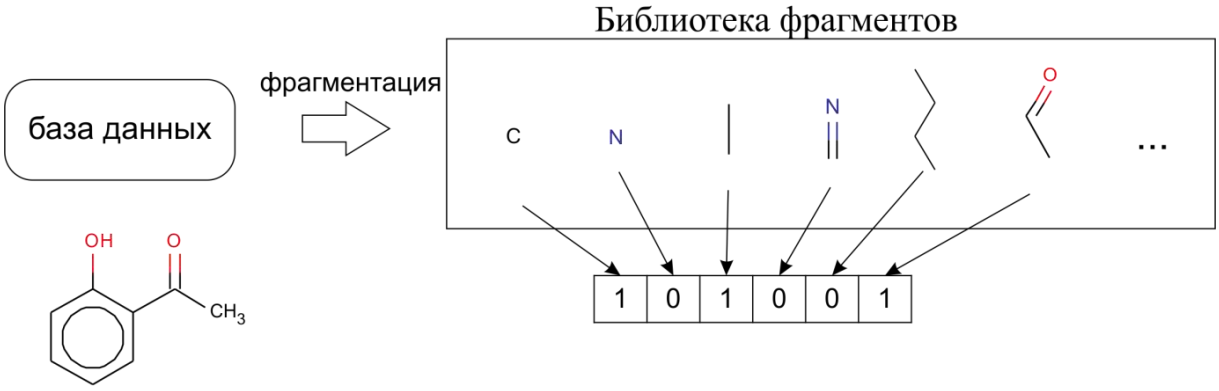


Рис. 15. Создание молекулярного отпечатка

Далее строится список всех полученных таким образом фрагментов, который и используется для создания молекулярного отпечатка (рис. 15). Битовая строка каждой молекулы определяется так же, как и в случае структурных ключей: наличие фрагмента дает 1 в соответствующем положении битовой строки, его отсутствие – 0.

Число идентичных фрагментов обычно не учитывается: если в молекуле присутствует несколько фрагментов данного типа, то активируется один единственный бит.

Молекулярные отпечатки лишены недостатка, характерного для структурных ключей – зависимости от заранее определенной библиотеки фрагментов. Они рассчитываются для конкретной базы данных и даже без специальной оптимизации обеспечивают хорошую структурную дискриминацию молекул. С другой стороны, длина битовой строки зависит от количества неидентичных фрагментов, на которые разбиваются все молекулы в базе данных (их могут быть тысячи), а количество единиц в ней для данного соединения определяется числом таких фрагментов в данном соединении (их обычно не больше 100). Из-за очень низкой доли единиц в битовой строке, формируемой для базы данных, содержащей разнородные соединения, эффективность их использования крайне низка. Такие битовые строки обладают низкой информативностью, поскольку содержат главным образом биты, соответствующие фрагментам, которые встречаются либо в одном соединении, либо почти у всех молекул (например, C, O и C-C). Кроме того, каждый раз, добавляя новую молекулу в базу данных, необходимо сравнивать все присутствующие в ней фрагменты с полным списком фрагментов для всей базы с целью определения положения соответствующих им бит в строке. По этим причинам молекулярные отпечатки в описанном выше виде никогда не используются. На практике все-

гда применяется один или несколько способов повышения эффективности работы с ними и сжатия данных в битовой строке: хеширование (англ. *hashing*), «складывание» (англ. *folding*)[17], создание молекулярных отпечатков варьируемой длины [45].

Хеш-ключи

Хеширование (иногда «хэширование», англ. *hashing*) – это преобразование объекта данных, обладающего сложной структурой и переменной длиной (например, представления молекулы), в целое число (в битовую строку фиксированной длины). В этом случае объект называется *хеш-ключем* (англ. *hash-key*), результирующее число – *хеш-кодом* (англ. *hash code*), а само преобразование – *хеш-функцией* (англ. *hash function*). Если хеш-код используется для определения адреса ячейки памяти, в которых располагаются хеш-ключ и связанные с ним данные в компьютере, то получается, что хеш-функции по значению хеш-ключа определяют их адрес и тем самым обеспечивают быстрый доступ к этой информации. Впервые поиск с использованием хеширования был использован Г.П. Луном из IBM в 1953 году, хотя само слово «хеш» появилось только в 1967 году [46].

Хеширование позволяет существенно ускорить работу с базами данных. Если необходимо проверить, является ли добавляемая в базу информация (например, структура молекулы) уникальной, надо осуществить ее поиск в базе данных. Если это достаточно крупная база данных – например, содержащая миллион соединений – проведение такого поиска может потребовать осуществить операцию сравнения данного соединения с миллионом других из базы данных, что в вычислительном плане крайне неэффективно. Если же все соединения хранятся в базе данных в соответствии с хеш-кодами, то, рассчитав для любого соединения хеш-код, можно сразу узнать, есть ли в базе данных о нем информация и где она хранится. Если хеш-код вычисляется исходя из структуры химического соединения, то он также позволяет проводить ультрабыстрый поиск по структуре соединения (рис. 16). По этой причине хеширование широко используются в хемоинформатике.

---

[17] В методе «складывания» молекулярных отпечатков пальцев это делается разделением битовой строки на несколько отрезков, которые потом «накладываются» друг на друга с помощью логической операции ИЛИ.

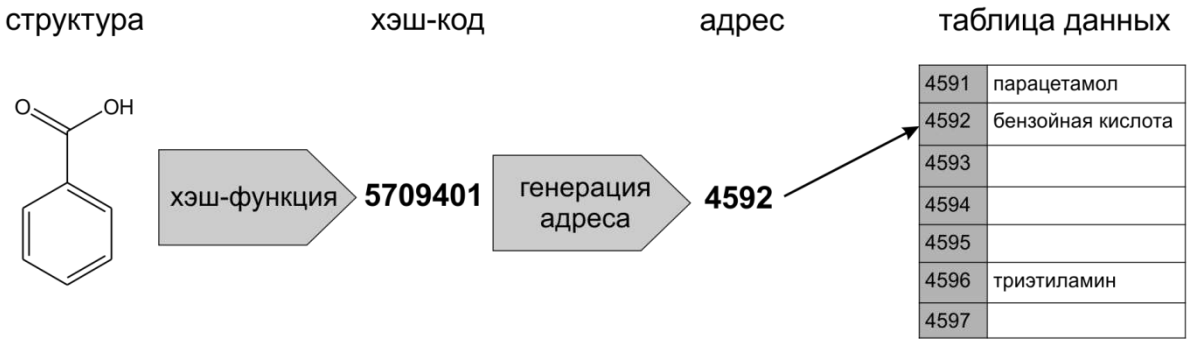


Рис. 16. Использование хеширования для поиска в базах данных

Задачей хеширования является получение числа определенной длины и при этом такого, что даже малые вариации в хеш-ключе (структуре химического соединения) приводят к заметному изменению хеш-кода. Если это условие не будет выполняться, то возможно, что хеш-коды очень похожих, но разных молекул совпадут. Вообще говоря, всегда существует вероятность того, что две молекулы, даже совершенно различные, будут иметь одинаковый хеш-код. Причиной тому является ограниченная длина хеш-кода для хранения молекул, тогда как число химических соединений практически бесконечно. Подобные случаи, так называемые коллизии, легко выявляются с помощью последующего непосредственного сопоставления молекул, и для битовых строк достаточно большой длины встречаются весьма редко[18]. Увеличение длины хеш-кода уменьшает вероятность коллизии. Поскольку хеш-код для двух близких молекул может сильно отличаться, но, с другой стороны, он однозначно определяется структурой молекулы, то говорят, что молекула служит затравкой для псевдо-случайного генератора чисел. По этой причине хеширование также называют генерацией псевдо-случайных чисел.

Хеш-коды представлены в компьютере как битовые строки, хотя для человека легче воспринимаются числа и буквы, поэтому зачастую

---

[18] Вероятность коллизии для полностью случайного хеша определяется уравнением:

$$P_E(n) = 1 - \prod_{i=1}^{n}\left(1 - \frac{i-1}{2^E}\right),$$

где E – число бит в строке, n – число молекул в базе данных. Для 64-битной строки и базы, содержащей 1 000 000 соединений, она равна 0.000002%, для 100 000 000 молекул она равна 0.03%, а для 32 битной строки и 100 000 000 молекул – практически 100%.

двоичные хеш-ключи переводят в легко читаемые десятеричные числа или комбинации букв и цифр. Например, каждой комбинации из четырех бит можно поставить в соответствие десятеричное или шестнадцатеричное число (в нем, кроме цифр от 0 до 9 есть 6 букв от A до F). Разбив хеш-код на 8 бит, можно его преобразовать в ASCII-символы (латинские буквы, цифры и знаки препинания). Поступают и обратным способом – генерируя десятеричное число или набор букв в качестве хеш-кода, которые компьютер интерпретирует как битовые строки.

Наиболее известным примером хеш-кодов является строка InChIKey длиной 27 буквенных символов, которая получается в результате хеширования из линейного представления InChI; выступающего в роли хеш-ключа. От свойств хеш-кодов в значительной мере зависит скорость обращения к базе данных. Поэтому алгоритм хеширования часто является ноу-хау компании и редко полностью документируются. Описан алгоритм, предложенный И. Гаштайгером и В.-Д. Иленфельдтом [47] и использованный в программах WODCA [48] и CACTVS [44]. Описан также в общих чертах первый алгоритм хеширования структуры в 64-битную строку (так называемый ACMF – *augmented connectivity molecular formula*), который введен в пользование в CAS Registry в 1979 году [49].

Хешированные молекулярные отпечатки

Хешированные молекулярные отпечатки (англ. *hashed fingerprints*) позволяют создать более сжатое и информационно-насыщенное представление за счет того, что используются битовые строки фиксированной длины с числом бит, значительно меньшим, чем общее число кодируемых фрагментов. Вследствие этого одному и тому же биту в битовой строке может соответствовать несколько подструктурных фрагментов. Кроме того, один фрагмент активирует несколько бит в разных положениях строки (рис. 17). Это решает проблему неэффективности использования битовой строки обычными структурными ключами и нехешированными молекулярными отпечатками из-за низкой доли единиц в ней. Хешированные молекулярные отпечатки были разработаны в компании Daylight Chemical Information Systems, Inc[19].

[19]Подробнее с хешированными молекулярными отпечатками можно познакомиться на сайте компании Daylight Inc. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html

Алгоритм формирования хешированных молекулярных отпечатков работает следующим образом. Для всех молекул из базы данных формируются списки присутствующих в них подструктурных фрагментов определенного типа, как в случае обычных молекулярных отпечатков. Далее для каждого фрагмента рассчитывается его хеш-код. Из хеш-кода вычисляется несколько адресов в битовой строке, которые соответствуют данному фрагменту (в этом случае хеш-код играет роль «затравки» для генератора псевдослучайных чисел – адресов в битовой строке). Таким образом, один фрагмент активирует несколько бит в строке (рис. 17 а). Поскольку битовая строка существенно короче, чем общее число фрагментов, то возникает достаточно много коллизий адресации – два разных фрагмента будут активировать один и тот же бит (рис. 17 б).

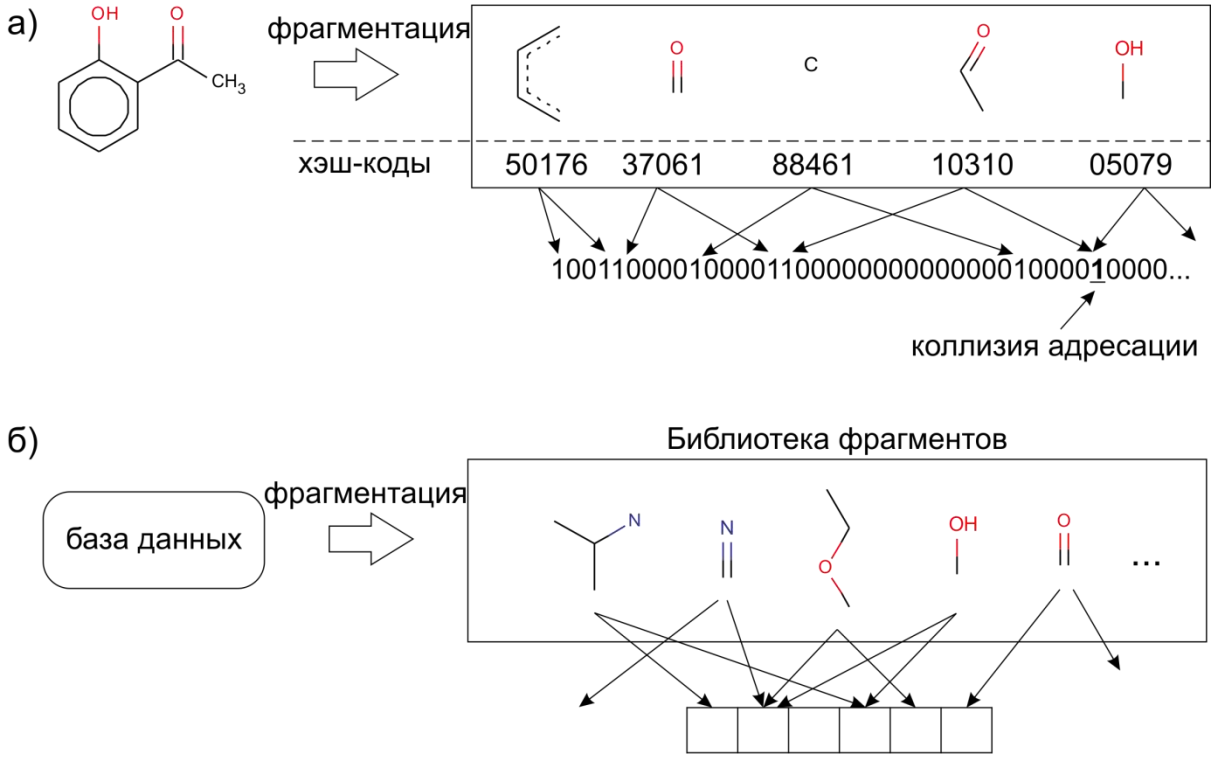


Рис. 17. (а) Создание хешированных молекулярных отпечатков, (б) биты в полученной битовой строке для всех соединений могут соответствовать многим фрагментам

Если бы каждый фрагмент активировал только один бит, то коллизия двух совершенно различных фрагментов, например C=O и C-O, приводила бы к тому, что две молекулы, имеющие все одинаковые фрагменты кроме этих, имели бы совершенно одинаковые молекулярные отпечатки – и в этом случае невозможно было бы отличить две разные молекулы (рис. 18 а). Если длина строки равна N бит, то вероятность того, что два фрагмента активируют один и тот же бит, равна

1/N (при N=512, вероятность коллизии бит 0.2%). Если каждый из фрагментов активирует не один, а два бита, то вероятность того, что эти два бита совпадут, уже равна 1/N(N-1) (при N=512, вероятность коллизии бит 0.0004%), то есть гораздо меньше. Таким образом, вероятность того, что два фрагмента активируют одну и ту же комбинацию бит, если каждый из них активирует два бита, существенно уменьшается (рис. 18 б). Это позволяет различать молекулы даже при наличии большого числа битовых коллизий.

Таким образом, хешированные молекулярные отпечатки определяются тремя параметрами – длиной битовой строки, размером фрагментов, и количеством бит, активируемых данным фрагментом[20]. Эти параметры требуют подбора при создании базы данных на основании ее размеров, поставленных задач и возможностей.

Увеличение длины строки приводит:

- к возможности более детально представлять структуру молекулы;
- уменьшению эффективности использования молекулярных отпечатков (меньшая плотность единиц – доля единиц в строке, в англ. используется также термин «темнота», *darkness*);
- увеличению требуемого пространства на хранение отпечатков;
- увеличение времени поиска по сходству (замедление работы с базой).

[20] В молекулярных отпечатках варьируемой длины величина битовой строки может определяться «плотностью» бит в строке. Подробнее – на сайте http://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

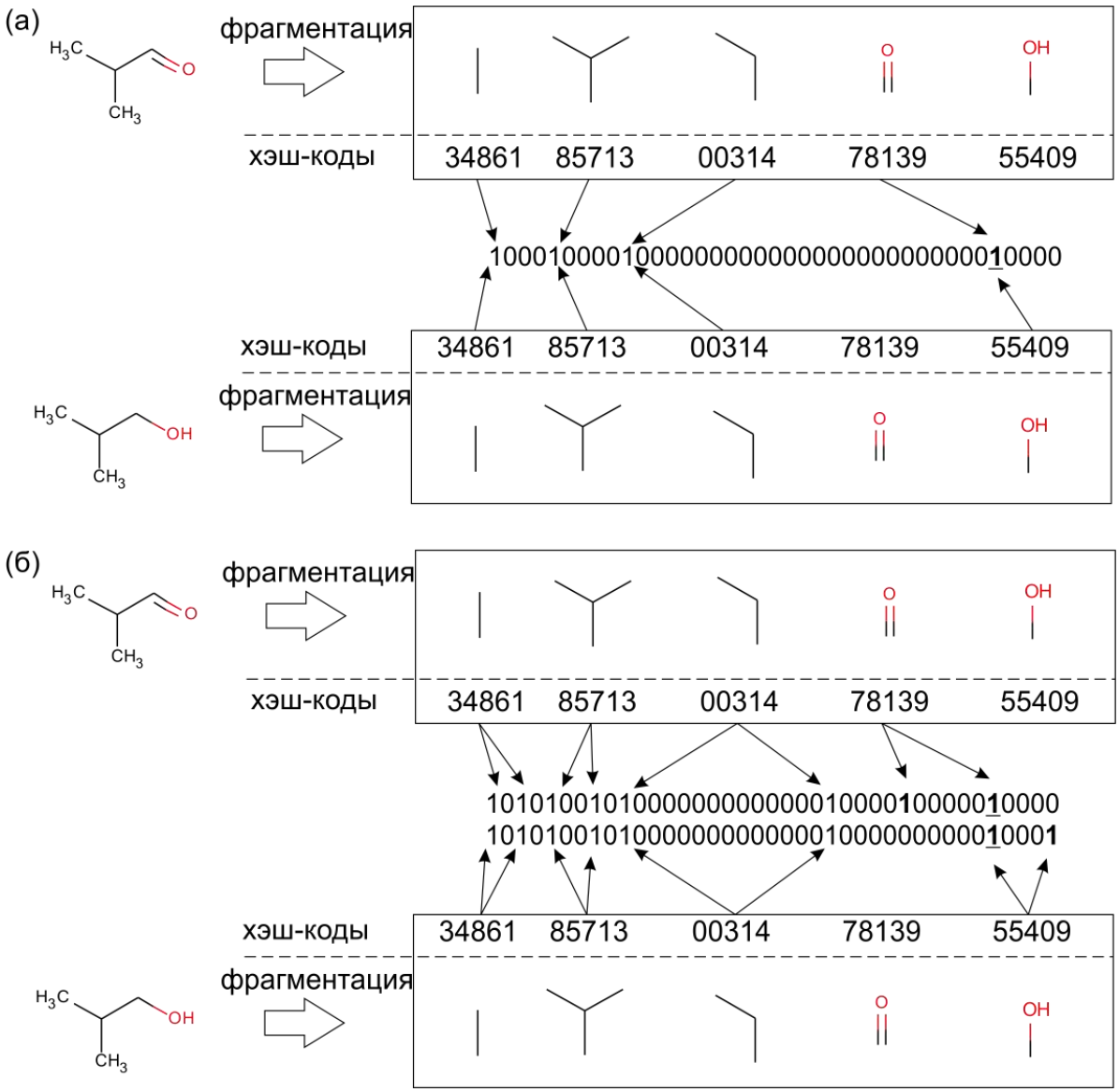


Рис. 18. Коллизия адресов битовой строки (а) ухудшает описание различных молекул, если каждый фрагмент активирует один бит. В случае (б), когда фрагмент активирует несколько бит, это слабо сказывается на дискриминации молекул

Компания ChemAxon рекомендует использовать в своих продуктах хешированные отпечатки длиной 512 бит (64 байта), поскольку они обеспечивают оптимальный баланс производительности и точности. В то же время, когда важна специфичность представления (при осуществлении поиска по сходству, кластеризации, при виртуальном скрининге по сходству), лучшим выбором будет использование строки длиной не менее 1000 бит. Увеличение максимально допустимых размеров фрагмента приводит к следующим последствиям:

- становится лучше представленной структурная информация о молекуле;

- из-за увеличения количества фрагментов добавление новой молекулы в базу замедляется;
- увеличивается количество фрагментов, что в свою очередь увеличивает плотность единиц в строке (при слишком большом их числе – обычно более 67-80% от общего числа бит – приводит к ухудшению дискриминации молекул, а при малом, менее 40% – к улучшению);
- растет число битовых коллизий;
- если количество коллизий не слишком велико, то увеличивается количество информации о структуре в битовой строке, что улучшает эффективность поиска.

Поиск по подструктуре обычно ведется с использованием фрагментов длиной до 5-6 атомов. Для виртуального скрининга рекомендуется использовать фрагменты длиной до 7 атомов. При увеличении длины больше 8 атомов дальнейшее улучшение обычно не наблюдается.

Увеличение количества бит, активируемых одним фрагментом, оказывает следующее влияние:

- увеличивается плотность единиц в строке с последствиями, описанными выше;
- улучшается качество кодирования структурной информации в битовую строку;
- увеличивается количество коллизий;
- структурная информация о молекуле передается лучше (близкие молекулы лучше дискриминируются).

Обычно один фрагмент активирует два бита. Дальнейшее увеличение не приводит к существенному улучшению результатов, так как в этом случае для компенсации увеличения числа коллизий и плотности единиц к строке требует увеличения длины строки или уменьшения числа фрагментов.

Хешированные молекулярные отпечатки обладают всеми преимуществами обычных молекулярных отпечатков, таких как независимость от соединений в базе данных и от типа запроса, обладая существенно большей компактностью представления. Они также на 20-40% более компактны, чем структурные ключи, без потери специфичности представления. Алгоритм хеширования не требует после фрагментации молекулы сравнения полученных частей с библиотекой фрагментов для выявления уникальности и определения положений активиру-

емых ими бит в строке. Получаемый из структуры фрагмента хеш-код сразу определяет адрес активируемого бита в строке. За счет этого скорость создания хешированных отпечаток очень велика.

С другой стороны, возможности их оптимизации весьма ограничены, поэтому в крупных базах данных обычно отдается предпочтение структурным ключам.

2.3.1.2. Векторное представление молекулы

Векторное представление является одним из наиболее широко используемых в хемоинформатике. **Вектором в информатике называют одномерный упорядоченный набор данных, имеющий фиксированную длину и служащий для хранения данных одного типа (обычно чисел), которые идентифицируются при помощи индекса (натурального либо неотрицательного целого числа)**. Иными словами, под вектором в информатике обычно подразумевается упорядоченный набор чисел, каждому элементу которого может быть присвоен порядковый номер.

Данный вид представления описывает каждую молекулу как совокупность числовых характеристик, называемых дескрипторами. С математической точки зрения, дескрипторы представляют собой *инварианты молекулярного графа* – характеристики графов, не зависящие от нумерации вершин и ребер. Эти характеристики могут отражать физико-химические свойства молекулы (например ее липофильность), топологические (например число связей), геометрические (например объем), квантово-химические (например энергия ВЗМО) характеристики и многие другие. Выбор дескрипторов для формирования вектора определяется исследователем в зависимости от задачи.

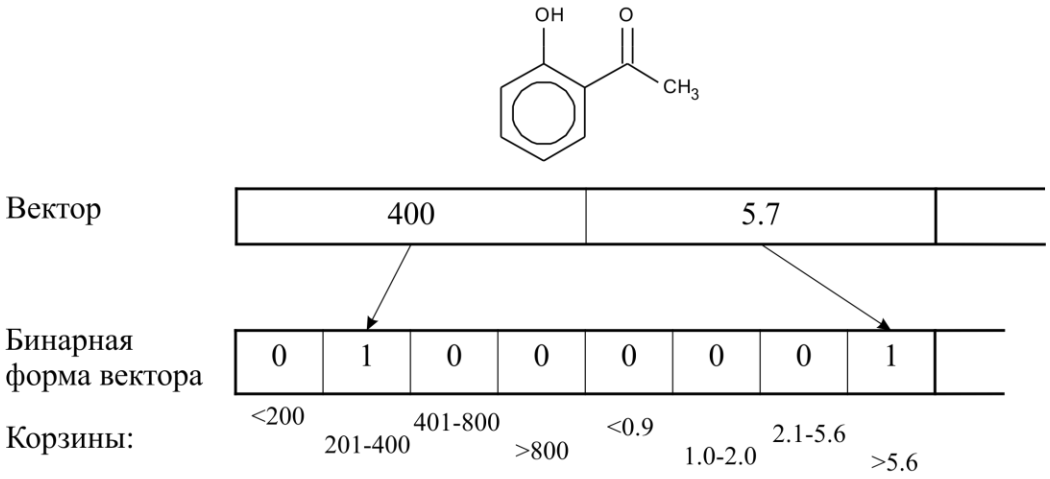


Рис. 19. Преобразование векторного представления в бинарную форму с использованием процедуры биннинга

Такой вид представления не является обратимым, поскольку в общем случае по набору дескрипторов невозможно восстановить структуру молекулы. Тем не менее, данное представление обладает одним важным преимуществом: вектор дескрипторов определяет координаты точки, соответствующей молекуле в построенном на N дескрипторах *N*-мерном пространстве. Тем самым векторное представление молекулы получает наглядную геометрическую интерпретацию. Предполагается, что близко расположенные в данном пространстве молекулы обладают сходными свойствами. В данном случае мерой близости молекул может являться эвклидово расстояние между соответствующими точками в этом пространстве. Это позволяет проводить поиск по сходству, строить классификационные и регрессионные модели с использованием таких представлений.

Векторное представление можно перевести в битовую форму с использованием процедуры *биннинга* (англ. *binning*). Для этого каждому числу векторного представления ставится в соответствие несколько бит в битовой строке, только один из которых «активирован» (т.е. принимает значение, равное единице). Каждому элементу (биту) полученной битовой строки соответствует определенный интервал значений, называемый корзиной (англ. *bin*). Бит, соответствующий данной корзине, активируется, если соответствующий элемент вектора попадает в интервал значений данной корзины (рис. 19). Интервалы значений каждой корзины обычно определяют так, чтоб обеспечить равномерный разброс единиц по корзинам. Бинарная форма записи вектора позволяет очень быстро вычислять меру сходства между мо-

лекулами благодаря использованию реализованных на аппаратном уровне операций с битами. Это может быть применено для поиска по сходству, кластеризации, построения классификационных и регрессионных моделей.

2.3.2. Матричное представление

Естественным математическим представлением графа является матрица. Матрица – это прямоугольная таблица чисел, которую можно рассматривать как совокупность строк и столбцов. Их число определяет размер матрицы. Матрицу также можно определить как функцию, которая по заданным номерам строки и столбца возвращает находящийся на их пересечении элемент. Поскольку матрица содержит две «координаты» – номера строки и столбца, с ее помощью можно, например, каждой паре вершин поставить в соответствие число, показывающее наличие ребра между ними. Такая матрица полностью описывает граф. Для описания молекулярных графов чаще всего используются квадратные матрицы, размер которых определяется числом атомов в молекуле. Молекула с ее атомами и связями тогда может быть представлена в матричной форме различными способами, в зависимости от того, какие характеристики пар атомов (их связанность, порядок связи между ними и др.) использованы для заполнения матрицы.

Обычно при построении матриц, описывающих молекулярные графы, атомы водорода в явном виде не учитывают, то есть используют т.н. *безводородные графы* (англ. *hydrogen suppressed graphs*). Это не приводит к потере информации, поскольку количество атомов водорода, связанных с атомом, всегда можно определить исходя из правил валентности.

Матрицы часто представляют информацию для описания молекулярного графа избыточно. Это происходит потому, что характеристика ребра (связи), соединяющего вершины (атомы) *i* и *j* указывается дважды – на пересечении *i*-й строки и *j*-го столбца, а также *i*-го столбца и *i*-й строки. Матрицы такого типа называются *избыточными* (англ. *redundant*). Избыточные матрицы симметричны относительно главной диагонали[21]. В неизбыточных матрицах характеристика ребра приво-

[21] Главная диагональ квадратной матрицы - диагональ, которая проходит через верхний левый и нижний правый углы. Матрицы, сим-

дится однократно, выше или ниже диагонали матрицы. Такую форму представления матрицы иногда называют треугольной.

Матричные представления молекул используются, как правило, для расчетов тех или иных характеристик молекул (например молекулярных дескрипторов), операций с молекулами (поиск по подструктуре, поиск пересечений). Для хранения информации в базах данных матричное представление, однако, применяется редко.

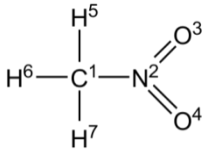
2.3.2.1. Матрица смежности

Матрицей смежности (англ. *adjacency matrix*) **называется квадратная матрица, чьи размеры определяются числом атомов в молекуле, а элементы определяют связность между атомами**. Каждый элемент матрицы равен нулю, если атомы с соответствующими номерами не связаны, или единице, если атомы связаны (рис. 20).

На диагоналях матрицы расположены нули. Матрица смежности является избыточной, число элементов в ней может быть уменьшено переходом к треугольной форме (рис. 20 в)[22]. Большой выигрыш в эффективности представления молекул без потери наглядности дает матрица смежности, в которой исключены из рассмотрения атомы водорода (рис. 20 г).

метричные относительно главной диагонали, называются симметричными.

[22] В данном случае под переходом к треугольной форме подразумевается организация такой формы размещения матрицы в памяти компьютера, чтобы в ней место занимали только элементы, находящиеся над или под главной диагональю. Не следует это путать с приведением матрицы к треугольному виду, которое осуществляется при решении систем линейных уравнений по методу Гаусса.



	1	2	3	4	5	6	7
1	0	1	0	0	1	1	1
2	1	0	1	1	0	0	0
3	0	1	0	0	0	0	0
4	0	1	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0

а)

	1	2	3	4	5	6	7
1		1			1	1	1
2	1		1	1			
3		1					
4		1					
5	1						
6	1						
7	1						

б)

	1	2	3	4	5	6	7
1		1			1	1	1
2			1	1			
3							
4							
5							
6							
7							

в)

	1	2	3	4
1		1		
2			1	1
3				
4				

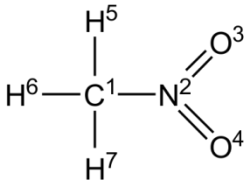
г)

Рис. 20. Избыточная матрица смежности нитрометана (а) и матрицы смежности с опущенными нулями: избыточная (б), неизбыточная (в), неизбыточная матрица для безводородного графа (г)

Размер матрицы смежности зависит от числа атомов, но не зависит от числа связей в системе. Она позволяет построить граф молекулы, но информации, которая в ней содержится, недостаточно для полного восстановления структуры, поскольку в матрице смежности отсутствует информация о типах атомов и порядках связей.

2.3.2.2. Матрица расстояний

Матрица расстояний (англ. *distance matrix*) является симметричной квадратной матрицей размерности NxN (число атомов на число атомов), элементами которой являются расстояния между соответствующими атомами в молекуле (рис. 21).



	1	2	3	4	5	6	7
1	0	1	2	2	1	1	1
2	1	0	1	1	2	2	2
3	2	1	0	2	3	3	3
4	2	1	2	0	3	3	3
5	1	2	3	3	0	2	2
6	1	2	3	3	2	0	2
7	1	2	3	3	2	2	0

а)

	1	2	3	4	5	6	7
1	0	1.480	2.305	2.305	1.097	1.097	1.097
2	1.480	0	1.218	1.218	2.103	2.103	2.103
3	2.305	1.218	0	2.170	3.151	2.442	2.789
4	2.305	1.218	2.170	0	2.442	3.152	2.788
5	1.097	2.103	3.151	2.442	0	1.830	1.802
6	1.097	2.103	2.442	3.152	1.830	0	1.802
7	1.097	2.103	2.788	2.788	1.802	1.802	0

б)

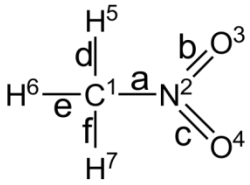
Рис. 21. Матрица расстояний молекулы нитрометана, построенная на (а) топологических расстояниях между атомами, (б) евклидовых расстояниях

Расстояния между атомами могут быть как евклидовыми расстояниями в трехмерном пространстве (выражаются в единицах длины, Å или нм), так и топологическими. Топологическое расстояние между атомами A и B – это минимальное количество связей, которое надо пройти по молекулярному графу от атома A к атому B.

Матрицы расстояний позволяют воссоздать расположение атомов в 3D пространстве или граф молекулы без указания типов атомов и порядков связей.

2.3.2.3. Матрица инцидентности

Матрица инцидентности (англ. *incidence matrix*) представляет собой прямоугольную матрицу, каждая строка которой соответствует определенной связи в молекуле, а столбец – атому. Пронумеруем отдельно все атомы и связи в молекуле. Если *i*-я связь соединяет *j*-й атом с каким-нибудь другим в молекуле, то в матрице инцидентности на пересечении *i*-й строки с *j*-м столбцом стоит единица, в противном случае – ноль (рис. 22 а). Эта матрица не является симметричной и избыточной. Число ее элементов может быть сокращено, если убрать нулевые значения (рис. 22 б) и не рассматривать атомы водорода (рис. 22 в).



	1	2	3	4	5	6	7
a	1	1	0	0	0	0	0
b	0	1	1	0	0	0	0
c	0	1	0	1	0	0	0
d	1	0	0	0	1	0	0
e	1	0	0	0	0	1	0
f	1	0	0	0	0	0	1

а)

	1	2	3	4	5	6	7
a	1	1					
b		1	1				
c		1		1			
d	1				1		
e	1					1	
f	1						1

б)

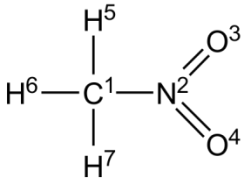
	1	2	3	4
a	1	1		
b		1	1	
c		1		1

в)

Рис. 22. Матрица инцидентности молекулы нитрометана (а) с указанием атомов водорода, (б) с опущенными нулями, (в) без указания водорода

2.3.2.4. Матрица связей

Матрица связей (англ. *bond matrix*) **представляет собой квадратную матрицу размером NxN (N-число атомов), элементами которой являются порядки связей между соответствующими атомами** (рис. 23).



	1	2	3	4	5	6	7
1	0	1	0	0	1	1	1
2	1	0	2	2	0	0	0
3	0	2	0	0	0	0	0
4	0	2	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0

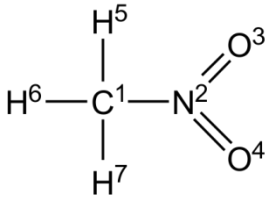
Рис. 23. Избыточная матрица связей молекулы нитрометана

Порядок связи для несвязанных между собой атомов формально полагается равным 0. Таким образом, элементами матрицы помимо единиц и нулей могут быть 2 (двойная связь), 3 (тройная связь) и т.д. Это представление также является избыточным и может быть сокращено удалением симметричной части или атомов водорода. Данная матрица предоставляет более полную информацию, чем матрица смежности, и может быть использована для полного восстановления вида молекулярного графа. Проблемой данного представления являет-

ся сложность представления ароматических связей, для которых нет конкретного значения порядка связи (условно можно полагать что 1.5 – но в данном случае требуется стандартизация).

2.3.2.5. Матрица связей-электронов

Матрица связей-электронов (англ. *bond-electron matrix или BE-matrix*) была предложена для использования в представлении реакций Уги и Дугунджи в 1973 году [50]. Она представляет собой матрицу связей, в которой на главную диагональ дополнительно помещены значения, равные удвоенному числу неподеленных пар электронов соответствующих атомов (рис. 24).



	1	2	3	4	5	6	7
1		1			1	1	1
2	1		2	2			
3		2	4				
4		2		4			
5	1						
6	1						
7	1						

Рис. 24. Матрица связей-электронов: избыточное представление с указанием атомов водорода

Поскольку в льюисовской модели полагается, что каждая связь образована парой электронов, то получается, что данное представление отражает распределение валентных электронов при образовании молекулы. Кроме того, в матрице связей-электронов сумма элементов по строке (или столбцу) равна числу валентных электронов соответствующего атома.

2.3.2.6. Другие матричные представления

Матричное представление молекул используется преимущественно для расчетов молекулярных дескрипторов. Существование большого числа всевозможных типов дескрипторов отчасти обусловлено широким разнообразием различных видов матричных представлений.

Матрица соседства ребер (англ. *edge adjacency matrix*) представляет собой квадратную симметричную матрицу, количество строк и столбцов которой определяется числом связей в молекуле. Элемент этой матрицы равен единице, если две связи содержат общий атом, и

нулю во остальных случаях. Атомы водорода обычно не учитываются при построении этого вида матриц. Матрицы соседства ребер используются для расчетов индексов соседства связей: спектральных моментов Эстрады [51-53], числа Платта, индекса Гордона-Скантлебери и др.

Для расчета различных топологических индексов используется также обратная матрица расстояний и обратная квадратичная матрица расстояний. В отличие от обычной матрицы расстояний, в них недиагональные элементы равны *(расстояние между атомами)*$^{-1}$ или *(расстояние между атомами)*$^{-2}$ соответственно. Диагональные же элементы в них равны нулю.

Матрица Бурдена [54] является модификацией матрицы смежности без учета атомов водорода, составленной по определенным правилам:

– диагональные элементы равны атомному номеру (порядковому номеру соответствующего химического элемента в таблице Менделеева);

– недиагональные элементы равны 0.1, 0.2, 0.3, 0.15 для одинарной, двойной, тройной и ароматической связи соответственно;

– для концевых связей к элементу матрицы добавляется 0.01;

– все остальные элементы имеют значения 0.001.

В оригинале они использовались для индексации соединений в базе данных, поскольку собственные значения [54] данных матриц были уникальными для всех соединений. Позднее на ее основе были созданы дескрипторы BCUT [55], широко используемые по настоящее время.

Матрица Лапласиана графа (матрица Кирхгофа) вычисляется как разница между диагональной матрицей порядков вершин графа (общего числа образуемых атомом связей без учета их порядков)[23] и матрицей смежности. Таким образом, каждый диагональный элемент в ней равен порядку вершины молекулярного графа, недиагональные элементы, соответствующие наличию связи между атомам, равны -1, а остальные – нулю. Матрица Лапласиана графа используется, в частности, для подсчета числа огибающих деревьев и расчета некоторых топологических индексов молекулярных графов [56], а также в спектральной теории графов.

Матрица обхода [57] похожа на матрицу топологических расстояний, но в ней топологические расстояния заменены на количество связей между вершинами графа по длиннейшему возможному пути. Мат-

---

[23] Диагональная матрица – матрица, содержащая ненулевые элементы только на главной диагонали.

рица расстояний-обхода является комбинацией матриц расстояний и обхода, верхний треугольник которой взят из матрицы обхода, а нижний – из матрицы расстояний.

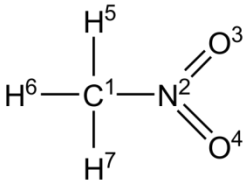
Существует множество других матричных представлений, которые используются для расчета разных видов молекулярных дескрипторов. Более детальный их обзор приведен в литературе [58].

2.3.3. Табличное представление

Таблицы являются более общим, чем матрицы, представлением, в котором каждый элемент данных помещен в ячейку, которая адресуется номером строки и столбца. В отличие от матриц, таблицы не обязательно должны иметь прямоугольную форму, и они могут содержать данные разного типа. Это дает возможность с их помощью представлять более сложные структуры данных.

Главным недостатком всех матричных представлений является квадратичный рост, с увеличением числа атомов, объема необходимой для их хранения компьютерной памяти. При этом большая часть этого объема расходуется нерационально, поскольку основная часть элементов этих матриц – нули. Для эффективного представления структур необходимо, сохраняя преимущества матричного представления, сделать так, чтобы объем необходимой памяти рос линейно с увеличением числа атомов в молекуле. Такое представление может быть основано на перечислении элементов химической структуры – атомов и связей. Именно таким образом устроены таблицы связности. Они состоят из двух списков – списка атомов и списка связей. В списке атомов перечисляются номера атомов, символы химических элементов, и, возможно, другие характеристики (например, формальный заряд на атоме, число присоединенных к атому атомов водорода, конфигурация хирального центра, тип изотопа, координаты атома и т.д.). Список связей содержит номера атомов, между которыми образованы связи, и типы связей. Списки атомов и списки связей связаны друг с другом через номера атомов. Типы связей могут быть как «реальными» и обозначать порядки связей, так и представлять собой нумерацию различных типов связей, которым нельзя приписать конкретного порядка (например, 5 может обозначать ароматическую связь, 6 – донорно-акцепторную и т.п.). Если каждая связь описывается один раз, то такие таблицы являются неизбыточными (рис. 25). Оба списка можно объединить в одну

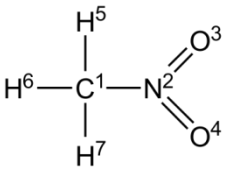
таблицу связанности, которая уже будет избыточной (рис. 26). В ней, как мы видим, каждая связь появляется дважды.



Список атомов	
1	C
2	N
3	O
4	O
5	H
6	H
7	H

Список связей		
Атом 1	Атом 2	Порядок связи
1	2	1
2	3	2
2	4	2
1	5	1
1	6	1
1	7	1

Рис. 25. Неизбыточная таблица связности, состоящая из двух списков



Атом	Символ	Атом 1	Порядок связи	Атом 2	Порядок связи	Атом 3	Порядок связи	Атом 4	Порядок связи
1	C	2	1	5	1	6	1	7	1
2	N	3	2	4	2				
3	O	2	2						
4	O	2	2						
5	H	1	1						
6	H	1	1						
7	H	1	1						

Рис. 26. Избыточная таблица связности, состоящая из одного списка

Таблицы связности были впервые предложены и введены в употребление Д. Глаком в 1961 году в компании DuPont [59]. С 1965 года они стали использоваться в Chemical Abstract Service (CAS) благодаря Гарри Моргану, разработавшему алгоритм уникальной нумерации атомов (алгоритм Моргана, см. выше). В 1979 году в CAS был создан алгоритм хеширования таблиц связности [49]. Генерируемые им коды ACMF являются первым примером хеш-кода, основанного на топологии молекулы. С 1980-х годов таблицы связности стали широко применяться при

создании баз данных. Большинство основных обменных форматов баз данных в тех или иных видах содержат таблицы связности.

Размер таблиц связности может быть дальше уменьшен исключением из рассмотрения атомов водорода, которые всегда могут быть восстановлены в соответствии с формальной валентностью атомов. Кроме того, функциональность таблицы связности может быть расширена добавлением дополнительных характеристик атомов (см. выше) и связей в существующие списки либо введением дополнительной информации (например, стереохимическую конфигурацию либо принадлежность атома стандартным остаткам при кодировании биологических макромолекул) при помощи специальных списков.

Таблицы связности полностью, однозначно и обратимо кодируют молекулярный граф молекулы. Хотя они не будут уникальными (поскольку зависят от нумерации вершин графа), применив известные алгоритмы канонизации, можно создать их каноническое представление [60].

Табличное представление молекул используются во многих программах как «внутренний» язык работы с молекулами, а также как основной формат хранения информации о структуре молекулы. Полное описание молекулы позволяет конвертировать данный вид представления практически в любой другой формат данных. Более того, основные форматы данных, используемые в хемоинформатике (PDB, MOL, RXN) являются расширением таблиц связанности и содержат их как основной элемент описания структуры. Конкретный вид таблиц связности может отличаться, однако основные принципы их формирования остаются неизменными.

2.4. СТРУКТУРЫ МАРКУША

Структуры Маркуша преимущественно используются для представления молекул в патентах [61]. Первая обобщенная патентная формула была использована в 1928 году Евгением Маркушем в патенте США №1506316 [62-64]. С тех пор обобщенные структуры Маркуша, называемые также структурными диаграммами Маркуша, обобщенными структурами или просто маркушами (рис. 27), используются крайне широко для этих целей.

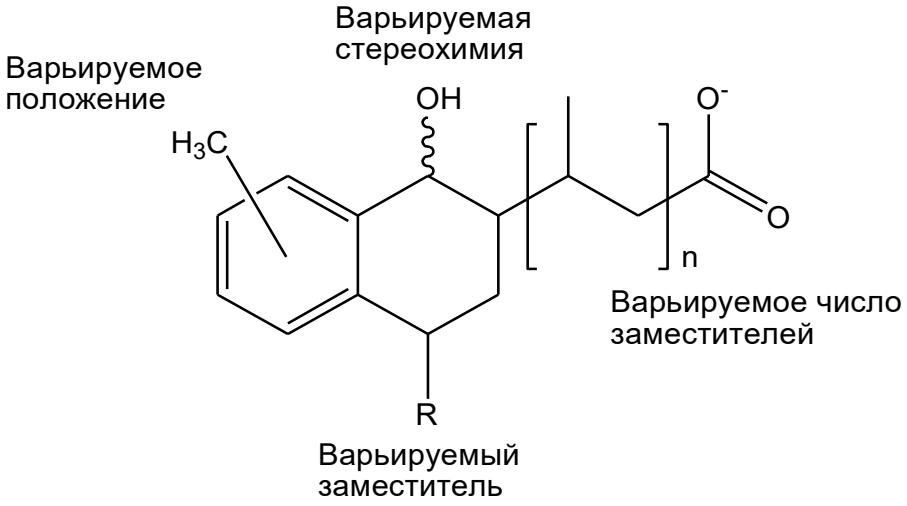


Рис. 27. Обобщенные структуры Маркуша и их возможности в указании семейства соединений

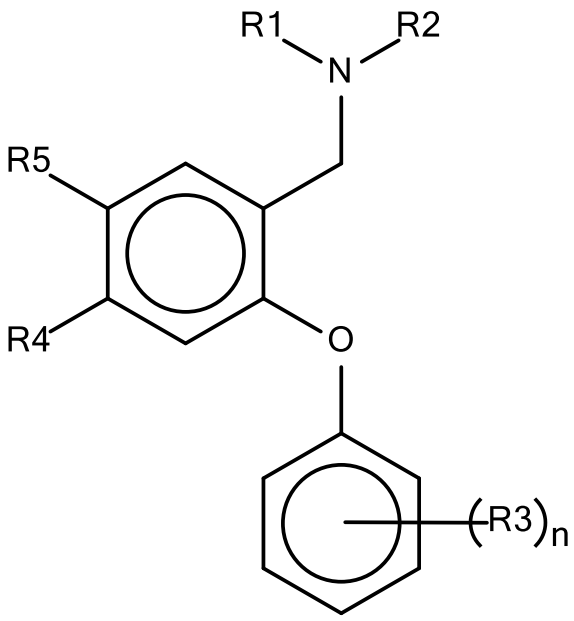
Первые попытки создания баз патентных данных с использованием достаточно примитивного поиска по структурам Маркуша были предприняты в начале 1960-х годов [65]. Тогда с ними работали точно так же, как и с обычными структурами. Начиная с 1980-х годов были разработаны две поисковые системы для структур Маркуша:

- французская Markush DARC (MDARC) [66], созданная компаниями Questel Orbit, Derwent Information и Французским патентным офисом (INPI), состоящая из нескольких подбаз (MPHARM, WPIM);
- американская система MARPAT [67, 68], разработанная в Chemical Abstract Service.

Обе эти базы получили существенное развитие благодаря разработкам группы профессора М. Линча в Университете Шеффилда [69]. В 1998 году базы данных из MDARC были объединены в одну систему MMS (*Merged Markush Service*). Сравнение MMS и MARPAT производится в работе Э. Беркса [70].

Структуры Маркуша – это специальный способ представления структур нескольких соединений на одном рисунке (рис. 27). В отличие от обычных структурных формул, они описывают целое семейство соединений, имеющих общий структурный мотив, и потому также называются обобщенными структурными диаграммами. Заместителями в структурах Маркуша могут быть не только конкретные группировки атомов (метил, этил, фенил), но также и их *обобщенные классы* (алкильные, арильные, гетероциклы). В отличие от других структурных представлений, спецификация варьируемых заместителей приво-

дится отдельно на рисунке или в тексте. Она может содержать достаточно сложное описание (рис. 28).



A compound of general formula (I), or a pharmaceutically acceptable salt thereof: wherein;

R1 and R2, which may be the same or different, are hydrogen, C1-C6-alkyl, $(CH_2)_m$(C3-C6-cycloalkyl) wherein m=0 1, 2 or 3, […];

each R3 is independently CF_3, OCF_3, C1-4-alkyl-thio or C1-C4-alkoxy; *n* is 1, 2 or 3; and

R4 and R5, which may be the same or different, are: A-X, wherein A= -CH=CH- or $-(CH_2)_p-$ where p is 0, 1 or 2; X is hydrogen, F, Cl, Br, I, […].

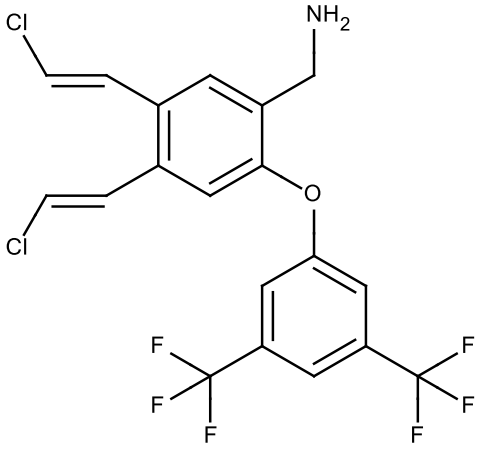
Рис. 28. Структура Маркуша из Заявки 1 патента США US-6448293-B1[24]. Взят отрывок из полного текста. Места сокращений обозначены как […]

Структуры Маркуша состоят из центрального фрагмента, называемого также *скелетом*[25] – на рис. 28 это замещенный дифениловый эфир. Данный центральный фрагмент может быть замещен ($-CH_2NR^1$, R^2, R^3, R^4, R^5). Положение заместителя может быть привязано как к конкретному атому (R^1, R^2, R^4, R^5), так и к любому атому в кольце (R^3). Заместители, чье число варьируется, обозначаются как $(X)_n$. Согласно приведенному на рис. 28 описанию патента, в кольце в разных положениях может быть до 3 заместителей R^3. При описании заместителя могут вводиться новые обобщенные заместители. Например, на Рис. 28 поясняется, что заместители R^4, R^5 имеют вид A-X, и указывается, что собой представляют новые обобщенные заместители A и X. Под приведенное на рис. 28

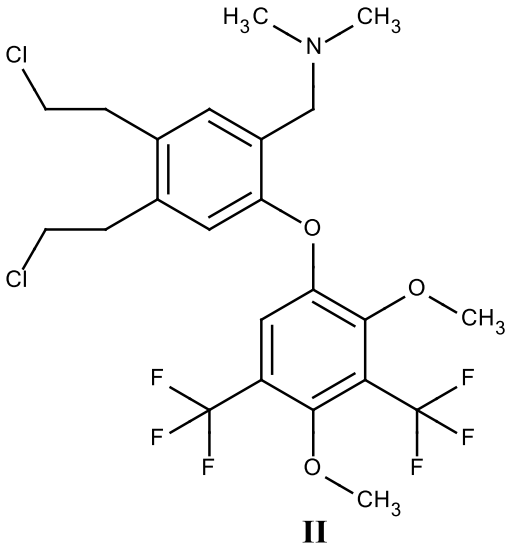
[24]Доступен по адресу http://patentsbase.com/items/US-6448293-B1-diphenyl-ether-compounds-useful-in-therapy

[25] В литературе также встречаются названия «скаффолд» и «фрэймворк» («каркас»), которые используются более широко в других разделах хемоинформатики (т.н. «скаффолд хоппинг»), и потому здесь не используются во избежание путаницы.

описание подходит, например, структура (**I**), но не подходит (**II**) потому, что содержит более чем 3 заместителя типа R^3.



I



II

Реализация компьютерного представления обобщенных структур является крайне важным для осуществления поиска в базах данных и оперирования маркушами (обмена информацией, визуализации, создания комбинаторных библиотек). В основе представления маркушей в базах данных лежат *расширенные таблицы связанностей* (ECTR, англ. *Extended Connection Table Representation*). С помощью ECTR представляются не только структуры, но и логические взаимосвязи между опциональными элементами структур Маркуша (называемых частичными структурами).

Для представления маркушей в виде расширенных таблиц связанностей (рис. 29) в обобщенной структуре выделяется центральный фрагмент и обобщенные заместители, которые могут присоединяться с родительской структурой по разным положениям. Заместители могут быть сами определены через младшие обобщенные заместители. Получается, что маркуш представляется в виде перевернутого дерева, корнем (родителем) которого служит центральный фрагмент, а связанные с ним обобщенные заместители (R^1, R^2,…) являются «стволами», приводящими к ветвям (также называемым «детьми») первого поколения, представленным возможными структурами заместителей. Например, если R^1 может быть -OH, $-NH_2$ и $-CH_3$, то на стволе, соответствующем R^1, существует три дочерние структуры одного поколения, относящиеся друг к другу логической операцией *ИЛИ*: -OH *ИЛИ* $-NH_2$ *ИЛИ* $-CH_3$ (рис. 29) Ветка-«ребенок» может являться «родителем» для младших веток. Например, если R^2 определена как группа $-CHR^3R^4$ (R^3 и R^4 могут быть независимо $-NH_2$ или $-CH_3$), то «ствол» R^3 представлен

одним ребенком -CHR^{3}R^4, от которого идут две ветки R^3 и R^4, на каждой из которых находится еще по две ветки (рис. 29). Таким образом, стволы относятся друг к другу операцией *И*, а «дети» одного поколения – операцией *ИЛИ*. Заметим, что структуры, зашифрованные в расширенных таблицах связанностей, не являются законченными химическими структурами (содержат опциональные точки присоединения или определены неявно: «алкил», «гетероцикл») и потому называются «частичными структурами». Частичные структуры могут быть определены явно (например, амино-группа) или неявно (например, алкенильные группы длиной от 4 до 6 атомов). Неявные структуры также специфицируются с использованием особого представления (т.н. описания обобщенных классов радикалов), определяющего структурные характеристики неявно указанных группировок с помощью определенных параметров: C – число атомов углерода, E – число двойных связей, Y – число тройных связей, RC – число колец, RA – число ароматических колец, Z – число гетероатомов.

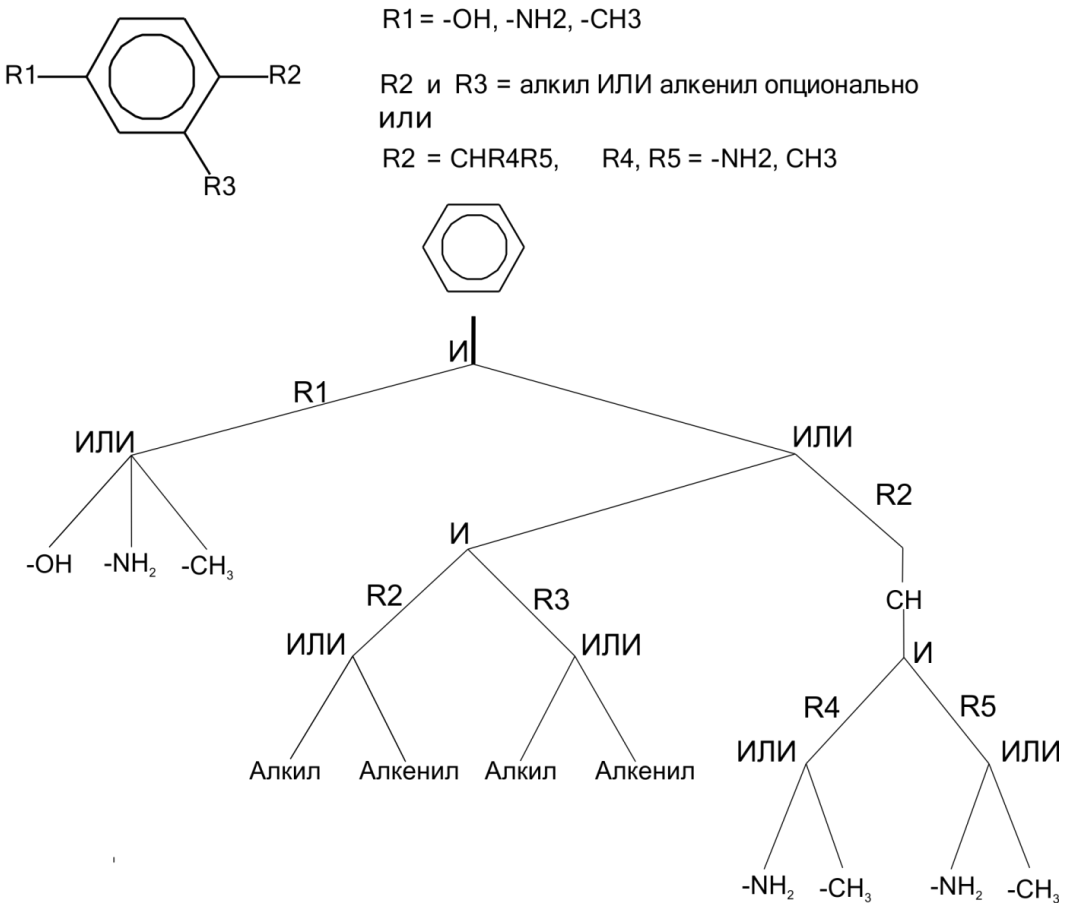


Рис. 29. Представление обобщенных структур в виде ECTR дерева. В верхней части рисунка приведена обобщенная структура замещенных аренов, в нижней части – соответствующее им ECTR дерево

2.5. ТРЕХМЕРНЫЕ ПРЕДСТАВЛЕНИЯ МОЛЕКУЛ

Очевидно, что сведений о том, как атомы в молекуле связаны друг с другом, зачастую недостаточно для понимания всего спектра явлений в химии. В ряде случаев принципиально важным является взаимное расположение атомов относительно друг друга в трехмерном пространстве, а не просто в виде молекулярного графа. Однако это сопряжено с проблемой необходимости учета возможных конформаций молекулы, которых, вообще говоря, бесконечно много, что существенно ограничивает использование трехмерных представлений.

Молекулы не являются структурно жесткими и могут совершать колебания и вращения относительно связей. Это приводит к тому, что атомы в молекуле находятся в постоянном движении друг относительно друга, реализуя непрерывный набор всевозможных «мгновенных» структур, называемых конформациями. Переходы между конформациями не сопровождаются изменениями связности атомов (топологии) молекулы. Энергетические барьеры при превращении одной конформации в другую могут быть легко преодолены за счет кинетической энергии молекул даже при обычной температуре[26]. Некоторые конформации являются более стабильными, и в таком состоянии молекулы проводят основную часть времени. Такие конформации, называемые конформерами, соответствуют локальным минимумам на поверхности потенциальной энергии системы. В соответствии с законом распределения Больцмана существует возможность того, что молекула может перейти в любой, даже самый нестабильный конформер, однако вероятность этого экспоненциально падает с ростом энергии данной конформации относительно наиболее стабильной. Энергия, требуемая на растяжение связей и изменения валентных углов, гораздо выше энергии, требуемой для изменений диэдральных углов. Поэтому переходы между конформерами сопровождаются главным образом изменением величин диэдральных углов.

Таким образом, для описания трехмерной структуры молекулы необходимо знание наиболее стабильного конформера (или набора таких конформеров). Если молекула характеризуется только одним кон-

---

[26] Более того, молекулы будут совершать колебания даже при температуре абсолютного нуля. Это явление называется «нулевыми колебаниями».

формером, то ставится вопрос, насколько представительным он является. Ответ на него зависит от изучаемого свойства. Физические свойства системы (например, спектры ИК и ЯМР) чаще всего будут определяться наиболее стабильным конформером или конформерами. Конформация же молекулы в активном центре фермента обычно отличается от наиболее стабильного в вакууме или в водном растворе конформера, причем иногда весьма существенно. Таким образом, поиск наиболее представительного конформера или набора конформеров является ключевой задачей при описании трехмерной структуры молекул и формирования на основе этого трехмерных представлений. Вопрос создания трехмерных представлений, исходя из химической структуры молекулы, будет рассмотрен в главе 2.7.2. «Конвертация представлений: 2D – 3D».

2.5.1. Координатные представления

Координатное представление является наиболее простым видом представления трехмерной структуры молекул. Рассматривая каждый атом как точку в евклидовом пространстве, мы можем задать его положение с помощью определенной системы координат. Для этого необходимо:

- выбрать тип системы координат;
- выбрать начало координат и направление осей;
- найти координаты атомов в выбранной системе координат.

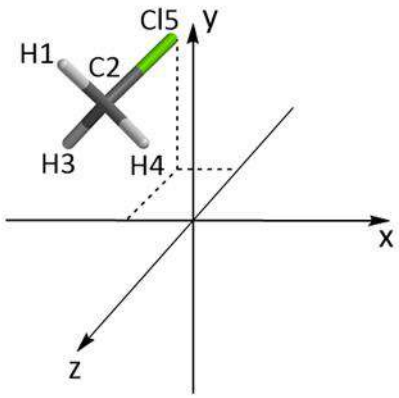
Координатные представления являются однозначными, но не уникальными, поскольку выбор начала координат и направления осей является произвольным. Обратимость же координатного представления, т.е. возможность точного восстановления структуры молекулы (включая связность атомов) по нему, далеко неочевидна. Хотя координатные представления задают только расположение атомов в пространстве, в идеальных случаях информация о связности атомов в молекуле и даже о порядках (типах) связей может быть восстановлена путем анализа межатомных расстояний. На практике же это может быть сопряжено с рядом проблем. Именно поэтому в файле, содержащем координатные представления, помимо координат атомов, часто приводится и таблица связности. В противном случае порядок связей будет рассчитываться программой на основании величин длин связей, гибридизации атомов, типов атомов, что для большого числа молекул может потребовать существенных затрат времени и дать ошибочные или неудобные для

анализа значения (например, двойная связь может быть ошибочно распознана как ароматическая и т.п.). Использование координатных представлений без указания типов связей и атомов может повлечь неадекватное понимание структуры молекулы программой, использующей данное представление для дальнейших расчетов, например, с использованием силовых полей. Информация о связности атомов не важна только для расчетов методами квантовой химии.

В молекулярном моделировании и хемоинформатике наиболее широко используются два типа координатных представлений: декартовы координаты и Z-матрицы. Эти представления легко могут быть переведены друг в друга.

2.5.1.1. Декартовы координаты

В данном представлении расположение каждого атома молекулы дается набором *x*, *y* и *z* – т.е. координат атомов в пространстве, которые равны расстоянию от начала координат до пересечения оси координат с проекцией положения атома на нее. Полученное представление является таблицей, содержащей в каждой строке идентификатор атома (иногда также номер атома и заряд ядра) и три координаты. Пример декартового представления молекулы приведен на рис. 30.



	x	y	z
H1	-1.8906	0.9874	-2.4444
C2	-1.8981	0.9874	-1.3522
H3	-2.4067	0.0935	-0.9847
H4	-2.4067	1.8813	-0.9847
Cl5	**-0.2322**	**0.9874**	**-0.7633**

Рис. 30. Декартова система координат и координаты хлорметана

Начало координат чаще всего помещают в центр масс или геометрический центр молекулы (точка, сумма расстояний от которой до всех атомов молекулы минимальна).

Формат декартового представления молекулы непосредственно в файле может отличаться в различных программах и может также содержать информацию о связности атомов и множестве других его ха-

рактеристик. Декартовое представление структуры молекулы с указанием таблицы связанности является основной формой представления молекул в хемоинформатике. Основные форматы файлов в хемоинформатике (PDB, MOL, RXN) используют такой вид представления трехмерной структуры молекулы.

2.5.1.2. Внутренние координаты (Z-матрицы)

Если декартовы координаты указывают положения атомов по отношению к внешнему объекту – осям координат, то *внутренние координаты* описывают расположение атомов относительно друг друга. Они используют в качестве координат длины связей, валентные и диэдральные углы. Стандартным представлением молекулы во внутренних координатах является Z-матрица (см. рис. 31).

Каждая строка Z-матрицы описывает один атом и начинается с его идентификатора, который обычно представляет собой комбинацию стандартного обозначения соответствующего химического элемента и порядкового номера атома в молекуле. Пусть порядковый номер текущего атома равен a_1. Далее в строке указывается длина связи a_1-a_2 (в ангстремах) и номер соседнего атома a_2, с которым текущий атом a_1 образует эту связь. Далее на этой же строке указывается валентный угол a_1-a_2-a_3 (в градусах) между связями a_1-a_2 и a_2-a_3 и номер атома a_3, удаленного от атома a_1 на две связи[27]. В последних двух столбцах на этой же строке указывается величина диэдрального угла a_1-a_2-a_3-a_4 (в градусах) и номер атома a_4, удаленного от атома a_1 на три связи[28].

C1						
C2	1.5	1				
H3	1.1	2	109.5	1		
H4	1.1	2	109.5	1	120	3
H5	1.1	2	109.5	1	-120	3
Br6	1.9	1	109.5	2	180	5
H7	1.1	1	109.5	2	60	5
H8	1.1	1	109.5	2	-60	5

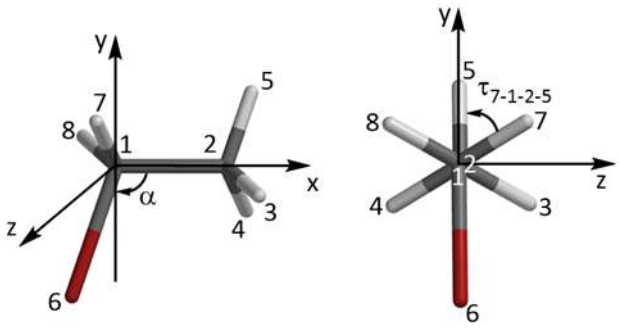


Рис. 31. Z-матрица бромэтана

[27] В органической химии такой атом называют геминальным.

[28] В органической химии такой атом называют вицинальным.

Проиллюстрируем алгоритм задания Z-матрицы на примере 1,2-дихлорэтана. Прежде всего, выбирается первый атом, положение которого полагается началом координат (им может быть любой атом молекулы). Его идентификатор помещается на первую строку Z-матрицы (атом C1 на рис. 31). Идентификатор второго атома (C2), связанного с первым, помещается на вторую строку, и указывается номер атома, с которым он непосредственно связан – то есть 1 – и длина связи (1.5 Å). На третьей строке указывается атом (атом C13 на Рис. 31), связанный с первым атомом, обозначается длина связи 1.7 Å и номер атома (1), с которым он связан. На этой же строке указывается величина угла между текущим атомом (с номером 3), его соседом (атом 1) и удаленным на две связи атомом 2 (величина угла 3-1-2 равна 109°), номер которого завершает эту строку. На четвертой строке приводится атом, связанный с любым из концевых атомов в цепочке 3-1-2 (атом H4 на рис. 31), длина его связи с соседним атомом (атом C2) и его номер, валентный угол с геминальным атомом (C1) и его номер, величины диэдрального угла с вицинальным атомом (Cl3) и его номер. Таким же образом строится вся остальная молекула, начиная с атомов, имеющих в качестве соседа тройку связанных атомов, уже обозначенных в таблице.

Z-матрицы, в отличие от декартовых координат, легко строить и анализировать человеку вручную. Кроме того, их преимуществом является то, что с их помощью легко вручную модифицировать структуру, тогда как при использовании декартовых координат всякий раз требуется пересчет координат всех атомов. Z-матрицы получили широкое распространение в квантовой химии и в методах, основанных на использовании силовых полей (молекулярной механике, динамике, докинге), тогда как в хемоинформатике обычно используют декартовые координаты, которые требуют указания меньшего числа параметров (не нужно указывать номера атомов) и с помощью которых легче визуализировать молекулы. Так же, как и декартовы координаты, Z-матрицы не описывают связности атомов во всей полноте. Для методов, которым такая информация необходима (например, молекулярная механика и динамика), следует дополнительно вносить ее в файл.

2.5.2. Молекулярные поверхности

Представление о молекуле как о наборе сферических атомов и цилиндрических связей, обладающих определенной жесткостью и направленностью, лишь частично соответствует действительности,

предоставляя тем не менее возможность описать ее «скелет». Скелет во многом определяет форму и свойства молекул. Т. Энгел и И. Гастайгер проводят интересную аналогию между человеком и молекулой [71]. Скелет человека во многом определяет его характеристики – форму тела, рост и даже функциональные особенности. Подобно тому, как мы на самом деле видим не скелет другого человека, а поверхность его тела, образуемую кожей, так и молекулу можно в определенной мере рассматривать как объект, имеющий собственную поверхность, которую можно «почувствовать» с использованием определенных физических методов и в ходе химических реакций.

Молекулы, как известно, не ведут себя в строгом соответствии с законами классической механики, и представление химических связей в виде «жестких стержней» либо «пружинок» является очень упрощенным. Более корректное описание молекул дает квантовая механика, с позиций которой они рассматриваются как набор ядер, расположенных определенным образом в пространстве, и электронов, распределенных между ними и образующих подобие «электронного облака». Это облако сконцентрировано сильнее всего вблизи ядер, при отдалении от которых плотность электронов резко падает и почти достигает нуля только на больших расстояниях от них. Притяжение отрицательно заряженных электронов, плотность которых повышена на линии связи благодаря положительной интерференции электронных волн, к положительно заряженным ядрам по обе стороны от нее обуславливает образование ковалентной химической связи.

Незаряженные молекулы при приближении друг к другу сначала слабо притягиваются вследствие так называемого дисперсионного взаимодействия, а потом начинают сильно отталкиваться из-за расталкивания электронов, причем тем сильнее, чем сильнее одна молекула «вжимается» в другую. Таким образом, молекулы ведут себя, как будто обладают мягкой поверхностью. Эти поверхности до определенного момента притягиваются друг к другу, но дальнейшему сближению «мешает» поверхность другой молекулы. Некоторые свойства молекулы (электростатический потенциал, заряд, гидрофобность) могут быть отображены на этой поверхности в виде цветной карты. Это отображение позволяет увидеть, что молекула неоднородна с разных сторон, и объяснить множество ее характеристик и свойств.

Вид молекулярных поверхностей зависит от трехмерных координат атомов, поэтому для их построения требуется на первом этапе

определить координаты всех атомов в молекуле из эксперимента или расчетов.

Молекулярные поверхности могут быть использованы как для визуального представления молекул, так и для расчета некоторых их характеристик (дескрипторов). Визуально поверхность молекулы можно представить в виде непрозрачной или прозрачной оболочки (англ. *solid* и *transparent*), набора разбросанных точек (англ. *dotted*) или сетки (англ. *mesh*). Для расчета свойств, поверхности представляются чаще всего в виде набора координат точек на ее поверхности – так называемой сетки (англ. *grid*), или набора функций, определяющих форму поверхности.

Анализ молекулярных поверхностей особенно полезен в тех случаях, когда решающую роль играют межмолекулярные взаимодействия, реакции и трехмерное распределение молекулярных свойств, например, при анализе и интерпретации реакционной способности или результатов докинга, создании лекарственных препаратов.

Молекулярные поверхности могут быть окрашены в соответствии со значениями рассчитанных на них локальных свойств (т.е. характеристик, которые могут быть рассчитаны в заданной точке пространства, т.н. *молекулярных полей*). Отображение каждого такого свойства на молекулярной поверхности дает новое представление молекулы. Способы расчета формы молекулярной поверхности могут существенно отличаться. Ниже мы опишем некоторые, наиболее распространенные виды молекулярных поверхностей.

2.5.2.1. Ван-дер-ваальсовая поверхность

Ван-дер-ваальсовая поверхность – это поверхность, образуемая пересекающимися центрированными на атомах сферами, чей радиус равен ван-дер-ваальсовым радиусам.

Ван-дер-ваальсовые радиусы определяются исходя из энергетически наиболее выгодного расстояния между двумя несвязанными атомами (рис. 32). Сближение молекул на расстояние, меньшее суммы их ван-дер-ваальсовых радиусов, требует затрат энергии. Поскольку ван-дер-ваальсовые радиусы атомов много больше ковалентных, сферы соседних атомов, чьи радиусы равны ван-дер-ваальсовым, пересекаются и проникают друг в друга. Внешняя часть образовавшейся фигуры и представляет собой ван-дер-ваальсовую поверхность молекулы. Объем, ограниченный ван-дер-ваальсовой поверхностью, называется ван-

дер-ваальсовым объемом молекулы. Существуют эффективные алгоритмы построения ван-дер-ваальсовой поверхности молекулы и расчета ограниченного ею объема.

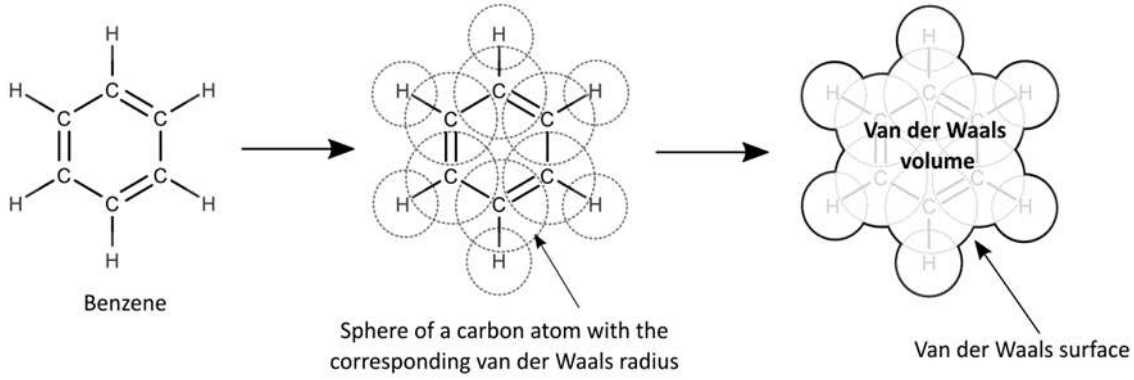


Рис. 32. Построение ван-дер-ваальсовой поверхности бензола

2.5.2.2. Поверхность Коннолли

При построении ван-дер-ваальсовой поверхности пересечение сфер создает острые и вогнутые участки на молекулярной поверхности, которые не могут быть доступны подходящим молекулам, и поэтому такие участки не годятся для представления формы молекул для предсказания возможности вступать в межмолекулярное взаимодействие.

В 1983 году М. Л. Коннолли предложил способ построения более гладких поверхностей [72, 73], которые он назвал поверхностями, доступными растворителю (англ. *solvent-accessible surface*). В настоящее время за ними закрепилось название «поверхности Коннолли», а название «поверхности, доступные растворителю», используется для другого типа поверхностей.

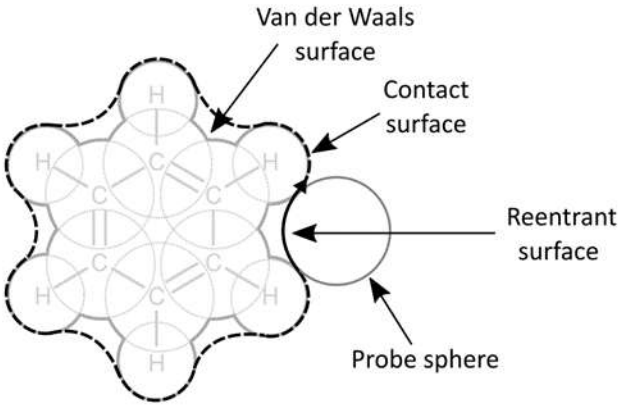


Рис. 33. Создание поверхности Конноли для бензола

Поверхность Коннолли есть поверхность, образованная сферической пробой (приближенно моделирующей растворитель), катящейся по ван-дер-ваальсовой поверхности (рис. 33). Радиус пробы обычно берется равным 1.4 Å, что соответствует эффективному радиусу молекулы воды. Поверхность Коннолли состоит из двух типов областей: выпуклых сегментов *контактных ван-дер-ваальсовых поверхностей*, непосредственно контактирующих с растворителем, и вогнутых сегментов *введенных поверхностей,* в которых растворитель касается ван-дер-ваальсовых сфер атомов молекулы двумя или более точками.

Объем, ограниченный поверхностью Коннолли, состоит из ван-дер-ваальсового объема молекулы и введенного объема. Этот суммарный объем молекула вытесняет при погружении в растворитель. Таким образом, поверхность Коннолли – это граничная поверхность, до которой может подойти растворитель. Существуют эффективные алгоритмы, основанные на гармоническом приближении, для расчета такого рода поверхностей [74].

2.5.2.3. Доступная растворителю поверхность

Доступная растворителю поверхность была предложена в 1971 году Б. Ли и Ф. Ричардсом для анализа взаимодействий пептид – растворитель, определяющих гидрофобность молекулы и форму протеина [75]. Чтобы построить эту поверхность, так же как и для поверхности Коннолли, используется «прокатывание» пробного шарика по ван-дер-ваальсовой поверхности.

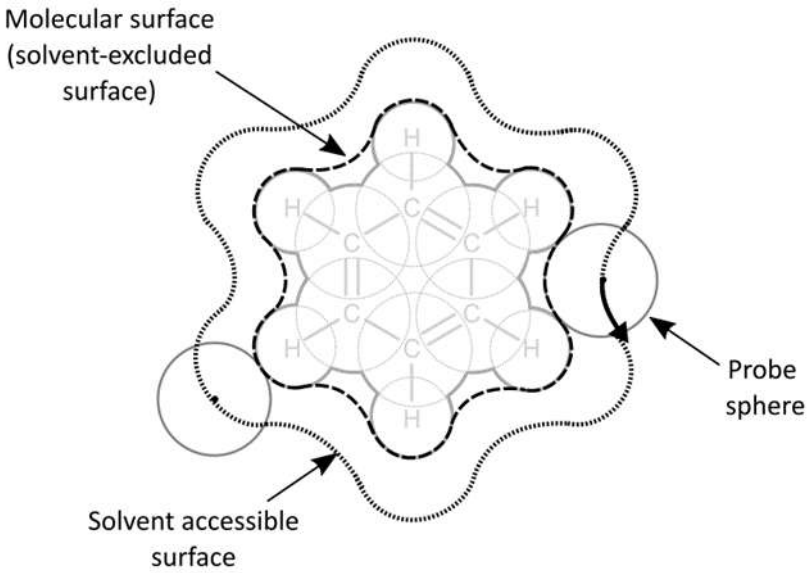


Рис. 34. Построение доступной растворителю поверхности молекулы бензола

Поверхность, определяемая положением центра шарика при таком «прокатывании», называется *доступной растворителю поверхностью* или *поверхностью Ли-Ричардса* (рис. 34). В то же время точки касания ван-дер-ваальсовой поверхности пробным шариком определяют поверхность Коннолли. Таким образом, **доступную растворителю поверхность можно определить как геометрическое место точек, образованных центрами молекул растворителя, окружающих данную молекулу** [76]. Доступная растворителю поверхность больше по площади, чем поверхность Коннолли, и объем, ограниченный ею, также больше.

2.5.2.4. Поверхность изоэлектронной плотности

Все указанные выше способы построения молекулярных поверхностей так или иначе касались использования ван-дер-ваальсовых поверхностей. Однако поверхности изоэлектронной плотности базируются на принципиально ином подходе – использовании характеристик распределения электронной плотности. Поверхность изоэлектронной плотности определяется как поверхность, в каждой точке которой электронная плотность равна заранее заданному постоянному значению. Было показано [76], что поверхности, образуемые точками, в которых значение электронной плотности равно 0.001 или 0.002 а.е.[29], очень близки к ван-дер-ваальсовым, и они приводят к значениям молекулярного объема, которые выводятся из физических экспериментов. Такие поверхности также называют «бейдеровскими поверхностями». Они отличаются от ван-дер-ваальсовых тем, что учитывают несферичность и влияние окружения атома, а также являются более гладкими.

Поскольку значения электронной плотности в каждой точке достаточно сложно вычислить теоретически (их можно рассчитать только с использованием методов квантовой химии), а также вывести из экспериментальных данных, этот вид представления поверхностей имеет ограниченное применение. Тем не менее такого рода поверхно-

[29] А.е. – атомные единицы. Для электронной плотности 1 а.е. электронной плотности это плотность распределения одного электрона в объеме 1 кубический бор (бор – атомная единица длины равная 0.529 Å).

сти часто используют для визуализации молекулярных форм, построенных на гауссовых функциях (например в программе vROCS).

2.5.3. Молекулярные формы

Молекулярная форма или объем представляют химическое соединение как область трехмерного физического пространства, занимаемого молекулой. Отличие молекулярной формы от поверхности в том, что формы моделируют то, что находится под (и над) молекулярной поверхностью, само «тело» молекулы.

Молекулярная форма описывается набором непрерывных функций, определенных во всех точках пространства. В соответствии с тем, как выбраны эти функции, можно выделить два основных подхода к представлению молекулярных форм:

1. *Гауссово представление*. В данном подходе форма молекулы описывается при помощи набора радиальных функций Гаусса[30], центрированных на атомах [77]. Поскольку значение радиальной функции Гаусса стремится к нулю лишь на бесконечном удалении от молекулы – такой тип представления не подразумевает наличия у молекулы четкой границы. Тем не менее интегрированием радиальных функций Гаусса по всему физическому пространству получаются конечные значения, которые можно связать с определенными свойствами молекул, например, с молекулярным объемом. Более того, интеграл произведения радиальных функций Гаусса, описывающих формы двух молекул, по всему физическому пространству также принимает конечное значение, которое описывает меру сходства между ними. Гауссовы функции, принадлежащие одному атому, можно «раскрашивать» и присваивать им свойства – H-донорность, гидрофобность и т.п. Дополнительным преимуществом применения радиальных функций Гаусса является то, что любой тип радиальной зависимости молекулярного поля может быть представлен как линейная комбинация гауссианов. Данный подход реализован в программе ROCS [78] и используется при проведении виртуального скрининга молекул на основе оценки сходства их молекулярных форм. Кроме того, Гауссово представление молекулярных форм лежит в основе метода непрерывных полей, позво-

[30] Радиальная функция Гаусса – это функция Гаусса (гауссиан), принимающая в качестве аргумента геометрическое расстояние до определенной точки в пространстве, называемой ее центром.

ляющего в сочетании с разнообразными методами машинного обучения прогнозировать биологическую активность органических соединений и осуществлять виртуальный скрининг молекул [79, 80].

2. *Разложение по сферическим гармоникам* [81]. В данном случае молекулярная форма представляется в виде набора центрированных в одной точке вещественных сферических гармоник. Сферические гармоники – это достаточно сложные функции, появляющиеся как решения определенных дифференциальных уравнений. В общем случае их значения являются комплексными (то есть содержат мнимую единицу, число *i*), однако можно найти их вещественные части. Существует бесконечное число таких функций, вид которых определяется числами *l* и *m* (m может принимать значения от -*l* до +*l*). Интеграл произведения двух таких функций при различных *l* или *m* равен нулю. Замкнутую поверхность любой формы с той или иной степенью точности можно представить как взвешенную сумму определенного количества сферических гармоник, чье число *l* изменяется от 0 до *L*. Наиболее известные химикам сферические гармоники – s-, p-, d- и f-атомные орбитали, чья форма определяется сферическими гармониками с l = 0 до 3. Математические свойства сферических гармоник позволяют легко находить перекрывания между молекулярными формами, сопоставлять их, находить общие фрагменты и др. Данный подход в приложении к молекулярным поверхностям реализован в программе ParaSurf [82].

2.6. СТАНДАРТНЫЕ ОБМЕННЫЕ ФОРМАТЫ ФАЙЛОВ

Существует большое количество компьютерных программ, которые работают со структурами молекул в химии. Одни программы созданы для того, чтобы рисовать структуры, другие работают с данными, получаемыми в ходе физических и физико-химических экспериментов (рентгеноструктурный анализ, хроматография, ЯМР-, УФ-, ИК- и масс-спектры), третьи – хранят данные (базы данных), четвертые – предсказывают свойства веществ. Все эти программы требуют того, чтобы информация была введена в них из файла или экспортирована из них в файл в том или ином формате. Отличия в функционировании программ приводят к тому, что форматы файлов для многих из них отличаются, и вывод одной программы может быть не совместимым со входом другой. Работа с данными, их анализ, извлечение из них информации (обработка данных) и дальнейшая выработка правил, зако-

нов, гипотез и теорий (то есть превращение информации в знание) требует обмена информацией между различными типами программного обеспечения: программами, базами данных, интернет-ресурсами. Вследствие этого ключевой задачей хемоинформатики является создание универсального, читаемого любой программой формата файлов для обмена химической информацией. Совершенно при этом не обязательно, чтобы этот формат использовался программой непосредственно – достаточно того, чтобы она могла конвертировать его в любой понимаемый только этой программой внутренний формат. Желательно также, чтобы универсальный формат обмена информации был текстовым. Исходя из этих требований, было разработано несколько форматов, которые в настоящее время являются стандартными в области хемоинформатики.

Предложенные в 1982 году компанией MDL коммерческие форматы сейчас приобрели огромную популярность в качестве универсального формата обмена данными и поддерживаются большей частью программ, используемых в хемоинформатике. Оригинальный формат представления молекул (MOL) компании MDL был расширен с тем, чтобы хранить информацию о свойствах молекулы (SDF-формат), реакциях (RXN-формат), а также иерархически организованные списки, содержащие дополнительную информацию (RDF-файл) [83]. Это позволяет хранить в одном файле как данные об индивидуальных молекулах, так и о реакциях. Другим распространенным форматом представления молекул является MOL2-формат компании Tripos Inc.

Для передачи информации о связности молекул часто используется линейное представление SMILES. В качестве одного из основных форматов представления трехмерной структуры биологических макромолекул, особенно белков и нуклеиновых кислот, используется формат PDB (англ. *Protein Data Bank*) [84, 85]. Наконец, универсальным форматом представления данных самого различного типа, от структуры молекул до спектральных данных, является основанный на стандарте XML язык химической «разметки» – формат CML.

2.6.1. Форматы MDL

Форматы данных MDL (англ. *Molecular Design Limited*), называемые также CTfiles (в настоящее время после ряда слияний правопреемником компании MDL является компания Accelrys), являются наиболее широко распространенными форматами файлов, используемыми в хемоинфор-

матике. Их можно назвать стандартными, поскольку они совместимы с большинством программ, применяемых в хемоинформатике. Одним из основных преимуществ этих форматов является то, что все они, несмотря на отличия в предназначении, связаны друг с другом. Второе существенное их преимущество заключается в возможности внесения в файлы информации о 2D и 3D структуре соединения, а также о практически любом свойстве молекул и реакций. Это позволяет использовать MDL-форматы для обмена информацией между разнообразными программами и системами управления базами данных.

Форматы данных MDL соотносятся между собой следующим образом. MOL-файл описывает структуру молекул. SDF-файл может содержать информацию о множестве молекул, структура каждой из которых задается в MOL-формате. Кроме того, SDF-формат позволяет хранить информацию о свойствах каждой молекулы (например, значения физико-химических свойств и виды проявляемой биологической активности). Это позволяет легко оперировать информацией, касающейся строения и свойств множества молекул, например, переносить данные о биологической активности тысячи молекул из базы данных в программу для анализа. RXN-файл описывает одну химическую реакцию в виде комбинации реагентов и продуктов, каждый из которых записывается в MOL-формате. RDF-формат, благодаря иерархической организации данных, позволяет, в частности, хранить информацию о свойствах реакций (например, температуре, давлении, выходе, селективности, ссылке на литературный источник).

Форматы MDL появились в эпоху использования перфокарт, что обусловило их некоторые особенности. Прежде всего, все содержащие информацию поля должны находиться в строке в строго фиксированных положениях, указанных в спецификации формата. Форматы всех чисел также должны быть строго фиксированы и соответствовать этой спецификации. По этой же причине несколько пробелов или нулей нельзя заменять на один пробел, удалять или перемещать. Кроме того, длина строки не должна быть больше 80 символов – рудимент из эпохи перфокарт. Подробная документация об этих форматах может быть получена бесплатно на сайте правопреемника компании MDL – компании Accelrys[31].

---

[31] http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/-no-fee.php

2.6.1.1. Структурный формат (MOL-формат)

MOL-формат позволяет достаточно детально описать молекулу. Он поддерживает возможность задания формальных зарядов атомов, изотопов и стереохимической конфигурации. Существует две версии формата, V2000 и V3000, которые несколько отличаются в представлении данных. Поскольку формат V3000 в настоящее время используются редко, его описание здесь не приводится. Подробную информацию по этой версии формата можно найти в официальной документации.

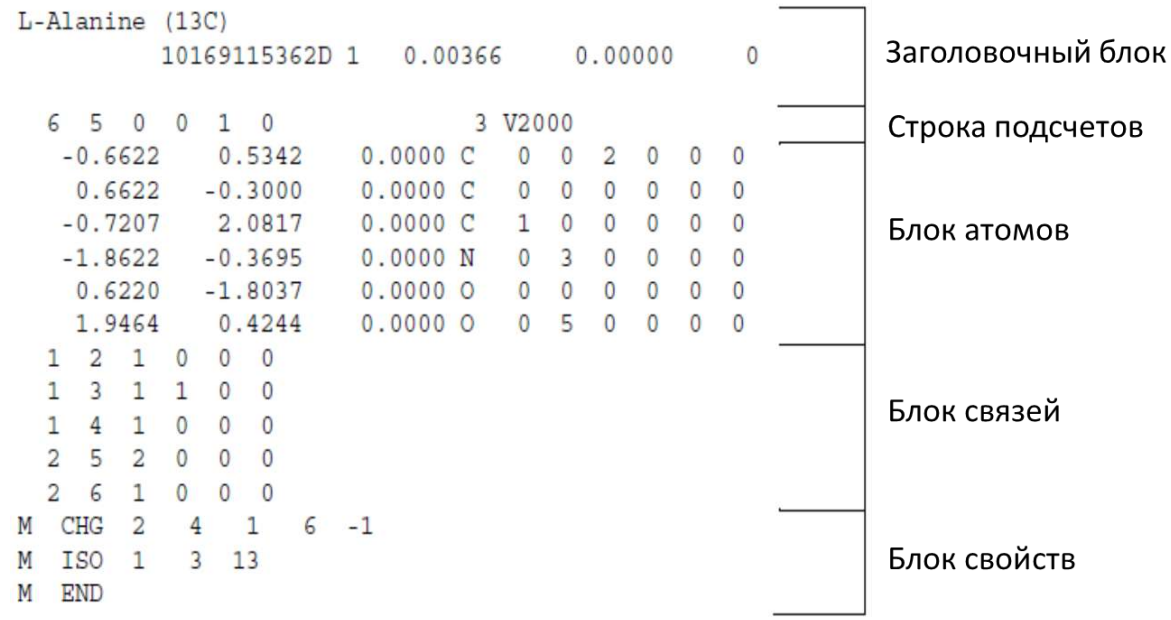


Рис. 35. MOL-файл, описывающий содержащую изотоп ^{13}C структуру L-аланина

MOL-файл состоит из нескольких блоков (рис. 35). *Заголовочный блок* состоит из трех строк: первая содержит название молекулы в произвольной форме и может быть пустой; вторая содержит информацию о молекуле (дата внесения в базу, шкалирующие факторы, энергия, номер) в специальной форме или может быть пустой; третья предназначена для комментариев и тоже может быть пустой. *Строка подсчетов* содержит информацию о числе атомов (первое поле), связей (второе поле), наличие хиральных атомов (пятое поле, 0 – нет, 1 – есть). Другие поля на этой строке заполняются редко. Последние 5 символов обозначают версию файла (V2000 или V3000).

Блок атомов содержит описание атомов. Каждый атом в молекуле описан на отдельной строке. В начале каждой из них приведены *x*, *y*, *z*-координаты соответствующего атома, причем они могут выражать

структуру молекулы как в 2D-представлении (тогда *z*-координаты равны нулю, а *x, y*-координаты предназначены для рисования структурной формулы на дисплее компьютера) или в 3D-представлении (тогда *x, y, z*-координаты обозначают положения атомов в физическом пространстве). Дальше идут обозначения символа атома, отличия массы изотопа от значения для наиболее стабильного (обозначает изотоп атома, от -9 до +9, 0 – если присутствует наиболее распространенный изотоп либо информация об изотопном составе не задана), заряд атома (от 0 до 9: 0 – не заряжен; 3 обозначает заряд +1; 5 обозначает заряд -1), атомная стереохимическая четность (для спецификации стереоизомерии, 0 означает «не указана», 1 означает «нечетная», 2 означает «четная»[32]), количество присоединенных атомов водорода (0 – не указано, число означает количество водородов + 1). Далее следуют несколько практически не используемых полей.

Блок связей следует сразу за блоком атомов и, по сути, является расширенной таблицей связей. Каждая строка в данном блоке соответствует одной химической связи и содержит:

- номера атомов (соответствующих нумерации в блоке атомов), между которыми образована данная связь;
- порядок связи (1-3 – ординарная, двойная и тройная, 4 – ароматическая, 7 – двойная или ароматическая, 8 – любая);
- стереоизомерию связей (см. документацию к CTfile) и некоторые другие редко используемые свойства связей.

В необязательном *блоке свойств* могут быть более детально указаны некоторые свойства (заряды, изотопы, радикалы и др.). При наличии данного блока указанные в блоках атомов и связей свойства не принимаются во внимание. Блок имеет следующий синтаксис[33]: M_[свойство (ISO-изотоп, CHG-заряд, RAD-радикал и др.)]_[количество атомов, для которых указаны свойства, не для всех свойств]_[описание свойства (например, указанием номера атома и его заряда)].

MOL-файл оканчивается строкой «M END».

---

[32] Правила стереохимической четности можно изучить по Гл. 2 книги: *Chemoinformatics. A textbook,* ed. J. Gasteiger and T. Engel. 2003, Weinheim: Wiley-VCH. 670 p.

[33] Здесь нижнее подчеркивание обозначает пробел.

2.6.1.2. Формат данных (SDF-формат)

Формат SDF (от англ. *Structure-Data File*) служит для хранения и обмена информацией о свойствах молекул. Структура каждой молекулы в SDF-файле записывается с использованием MOL-формата ([Molfile] на Рис. 36). Молекулы отделяются друг от друга строкой, содержащей символы «$$$$».

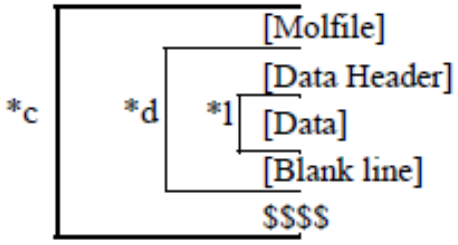


Рис. 36. Структура SDF-файла, *l – повторяется для каждой строки в данных (если требуется несколько строк), *d-повторяется для каждого данного (свойства), *c – повторяется для каждой молекулы.

Структура каждой молекулы может быть дополнена в SDF-файле списком ее свойств. Каждому свойству соответствует блок, состоящий из строки с названием свойства ([Data Header] на Рис. 36) и строки или нескольких строк с описанием свойства ([Data] на Рис. 36). В большинстве случаев строка с описанием свойства содержит его числовое значение, однако текстовые данные (например, название, SMILES или код соединения) допустимы. Название свойства задается с использованием следующего синтаксиса: > [номер молекулы (не обязательно, указан последними символами во второй строке MOL файла)] < [наименование свойства английскими буквами, избегая символов -, <, >, =, %, . и пробелов] <. На следующей строке или строках приводится описание свойства. После каждого свойства оставляется пустая строка.

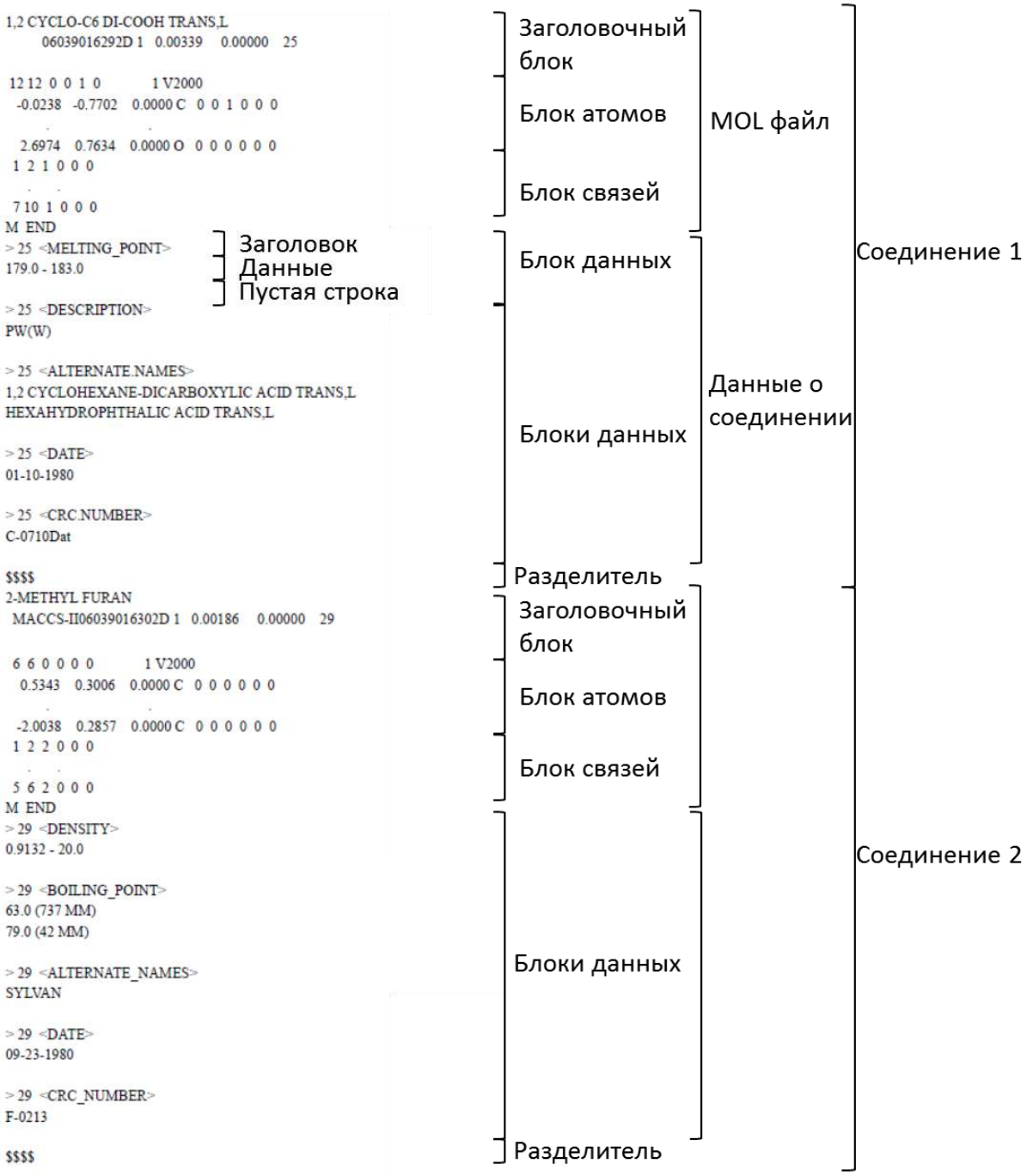


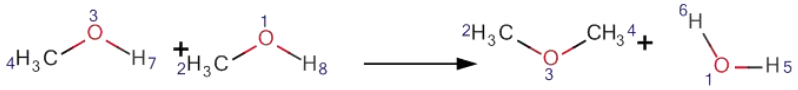
Рис. 37. Пример SDF-файла. Координаты атомов и описания связей частично пропущены

Знак $$$$ соответствует концу записи о данной молекуле. На рис. 37 приведен пример SDF-файла и расшифровка информации в нем.

2.6.1.3. Реакционный формат (RXN-формат)

Реакционный формат RXN служит для кодирования химических реакций путем указания их реагентов и продуктов. RXN-файл начинается со строки «$RXN», за которой следует заголовочный блок, содер-

жащий три строки: для названия реакции, для служебной информации и для комментариев. Далее следует строка, содержащая количество реагентов и количество продуктов реакции. Каждый реагент и продукт задаются последовательно в MOL-формате, причем сначала перечисляются реагенты, потом продукты. Описание каждой молекулы начинается строкой «$MOL».



```
$RXN

 Marvin      042001121925

 2  2
$MOL

 Mrv0541 04201219252D

 3  2  0  0  0  0            999 V2000
  -5.3036   -1.2375    0.0000 C   0  0  0  0  0  0  0  0  0  4  0  0
  -4.5891   -0.8250    0.0000 O   0  0  0  0  0  0  0  0  0  3  0  0
  -3.8746   -1.2375    0.0000 H   0  0  0  0  0  0  0  0  0  7  0  0
 1  2  1  0  0  0  0
 2  3  1  0  0  0  0
M  END
$MOL

 Mrv0541 04201219252D

 3  2  0  0  0  0            999 V2000
  -2.5634   -1.3259    0.0000 C   0  0  0  0  0  0  0  0  0  2  0  0
  -1.8489   -0.9134    0.0000 O   0  0  0  0  0  0  0  0  0  1  0  0
  -1.1345   -1.3259    0.0000 H   0  0  0  0  0  0  0  0  0  8  0  0
 1  2  1  0  0  0  0
 2  3  1  0  0  0  0
M  END
$MOL

 Mrv0541 04201219252D

 3  2  0  0  0  0            999 V2000
   2.7107   -1.2375    0.0000 O   0  0  0  0  0  0  0  0  0  3  0  0
   3.4252   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  4  0  0
   1.9962   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  2  0  0
 1  2  1  0  0  0  0
 1  3  1  0  0  0  0
M  END
$MOL

 Mrv0541 04201219252D

 3  2  0  0  0  0            999 V2000
   5.4509   -1.3554    0.0000 O   0  0  0  0  0  0  0  0  0  1  0  0
   6.2759   -1.3554    0.0000 H   0  0  0  0  0  0  0  0  0  5  0  0
   5.0384   -0.6409    0.0000 H   0  0  0  0  0  0  0  0  0  6  0  0
 1  2  1  0  0  0  0
 1  3  1  0  0  0  0
M  END
```

Рис. 38. Реакция образования диметилового эфира с указанием отображения атомов (обведено пунктиром) и вид соответствующего RXN-файла

Важным для задания реакций является отображения атомов исходных веществ на атомы продуктов, т.е. указание того, какой атом в

реагентах переходит в какой атом в продуктах реакции[34]. Для этого уникально нумеруется каждый атом, входящий в состав реагентов и продуктов реакции. В процессе этого определяется, в какой атом продукта превращается каждый атом реагента, и этим атомам присваиваются одинаковые номера. Тем самым задается отображения атомов.

Таким образом, каждый атом продуктов имеет свой «источник» в реагентах. Каждый атом продуктов может иметь только одно отображение в реагентах. Таким образом, создание отображения требует, чтобы в нем участвовало одинаковое число атомов в реагентах и продуктах реакции. В каждом из указанных MOL-файлов реагентов и продуктов в 14 столбце блока атомов указывается нумерация атомов, задающая их соответствие (отображение). Пример RXN-файла для реакции образования диметилового эфира приведен на рис. 38.

2.6.1.4. Формат реакций и данных (RDF-формат)

Формат реакций и данных RDF (от англ. *reaction-data file*) создан для того, чтобы хранить информацию о свойствах реакций и молекул (рис. 39).

RDF-файл содержит заголовочный блок, состоящий из двух строк: первая – «$RDFILE 1» (цифра «1» обозначает версию, других версий пока не существует) и вторая, которая начинается с ключевого слова «$DATM» и содержит информацию о дате создания файла (эта строка считается программой-комментарием). Файл может содержать описание и свойства множества реакций или соединений. Формат RDF, в принципе, можно использовать для хранения в одном файле как реакций и соединений, так и их свойств. Описанию каждой молекулы (задействованной в реакциях и не фигурирующей в RXN формате) должна предшествовать строка, содержащая одно из следующих слов: $MFMT $MIREG [внутренний номер] (внутренний номер – номер внутри вашего файла или базы данных), $MFMT $MEREG [внешний номер] (номер во внешней базе данных, например CAS номер), или просто $MFMT (без указания номера). Далее следует описание молекулы в MOL-формате. Аналогично описанию каждой из реакций должна предшествовать строка: $RFMT $RIREG [внутренний номер],

[34] Этот вопрос уже кратко рассматривался в главе 2.2.3.4. «Представление реакций (SMIRKS)» и будет обсуждаться далее в части «2.8. Представление реакций».

$RFMT $REREG [внешний номер], или $RFMT. Со следующей строки идет описание реакции в формате RXN.

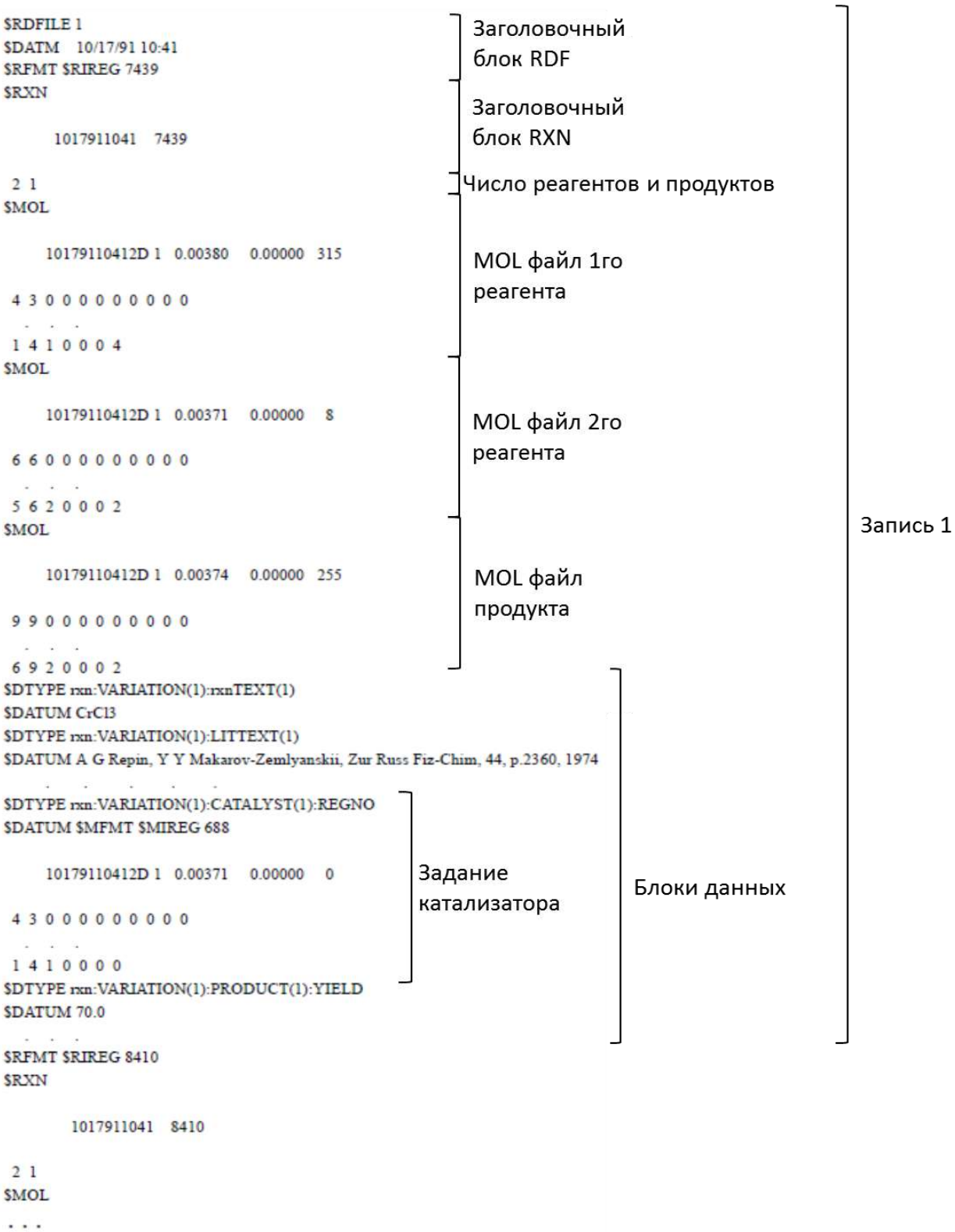


Рис. 39. Пример RDF-файла с расшифровкой

Описание свойств реакций или молекул следует за их спецификацией в RXN или MOL-формате. Как и в SDF-формате, каждое свой-

ство содержит его название на одной строке и описание на другой. В RDF-формате, однако, нет необходимости оставлять пустую строку между различными свойствами. Строка с названием свойства начинается с ключевого слова $DTYPE, после чего следует собственно название. Описание данных следует на отдельной строке, начинающейся с ключевого слова $DATUM, после которого следует описание свойства. Если описание свойства включает спецификацию химического соединения (например, при указании катализатора, растворителя), то на этой строке пишется $DATUM $MFMT, или $DATUM $MEREG [внешний номер соединения], или $DATUM $MIREG [внутренний номер соединения]. Сама молекула описывается со следующей строки в MOL-формате (см. рис. 39).

2.6.2. Молекулярный формат Sybyl mol2

Формат Sybyl mol2 – это емкий формат текстового файла представления молекул. Также его называют Tripos mol2 по названию компании-разработчика программы Sybyl, для которой этот формат был разработан. Данный формат позволяет описывать связность и 3D-структуру не только малых молекул, но также биологических макромолекул (белков, нуклеиновых кислот и полисахаридов), задавать фрагменты, типы атомов для методов молекулярной механики, периодические условия и многое другое. В отличие от форматов MDL, формат Sybyl не содержит ограничений на длину строки и не фиксирует точные положения полей в ней. Пустые строки, а также линии, заполненные пробелами или знаками табуляции (вместе называемые «белыми пространствами»), игнорируются и могут быть размещены в любом месте.

Строки, начинающиеся со знака #, считаются комментариями и игнорируются при чтении файла. Они могут присутствовать в любом количестве и в любом месте файла. На Рис. 40 в качестве комментария (который был автоматически создан программой) указано имя молекулы, автор и дата создания, а также информация о модификации файла.

Характеристики молекулы описываются в виде так называемых *записей* (англ. *data records*). Каждая запись имеет идентификатор типа записи, начинающийся с фразы @<TRIPOS>, и некоторое число полей данных, число которых в зависимости от типа записи может быть ограничено (для @<TRIPOS>MOLECULE – максимум 6 строк) или

сколь угодно большим (для @<TRIPOS>ATOM – не ограничено, по числу атомов). Формат полей данных зависит от типа записи.

```
#       Name: benzene
#       Creating user name: tom
#       Creation time: Wed Dec 28 00:18:30 1988

#       Modifying user name: tom
#       Modification time: Wed Dec 28 00:18:30 1988

@<TRIPOS>MOLECULE
benzene
12 12 1 0      0
SMALL
NO_CHARGES


@<TRIPOS>ATOM
1       C1      1.207   2.091   0.000   C.ar    1       BENZENE0.000
2       C2      2.414   1.394   0.000   C.ar    1       BENZENE0.000
3       C3      2.414   0.000   0.000   C.ar    1       BENZENE0.000
4       C4      1.207   -0.697  0.000   C.ar    1       BENZENE0.000
5       C5      0.000   0.000   0.000   C.ar    1       BENZENE0.000
6       C6      0.000   1.394   0.000   C.ar    1       BENZENE0.000
7       H1      1.207   3.175   0.000   H       1       BENZENE0.000
8       H2      3.353   1.936   0.000   H       1       BENZENE0.000
9       H3      3.353   -0.542  0.000   H       1       BENZENE0.000
10      H4      1.207   -1.781  0.000   H       1       BENZENE0.000
11      H5      -0.939  -0.542  0.000   H       1       BENZENE0.000
12      H6      -0.939  1.936   0.000   H       1       BENZENE0.000
@<TRIPOS>BOND
1       1       2       ar
2       1       6       ar
3       2       3       ar
4       3       4       ar
5       4       5       ar
6       5       6       ar
7       1       7       1
8       2       8       1
9       3       9       1
10      4       10      1
11      5       11      1
12      6       12      1
@<TRIPOS>SUBSTRUCTURE
1       BENZENE1        PERM    0       ****    ****    0       ROOT
```

Рис. 40. Молекула бензола, записанная в формате Sybyl mol2

Остановимся кратко на некоторых типах записей. @<TRIPOS>MOLECULE дает базовое описание молекулы и содержит:

- имя молекулы (1-я строка);

- число атомов, связей (если указаны), подструктур (если заданы) и некоторые другие характеристики (2-я строка);
- указание типа молекулы (SMALL, BIOPOLYMER, PROTEIN, NUCLEIC_ACID, SACCHARIDE – 3-я строка);
- типы зарядов атомов (NO_CHARGES- заряды атомов не указаны, существует большой выбор разных типов зарядов – 4-я строка);
- в 5-й и 6-й строке может содержаться служебная информация о молекуле, создаваемая программой и комментарий (на рис. 40 их нет).

Запись @<TRIPOS>ATOM задает описание атомов в составе молекулы. Каждая строка содержит номер атома, его уникальный идентификатор (составленный из обозначения химического элемента и порядкого номера атома данного типа), *x*, *y*, *z* – координаты, тип атома (например, C.ar – ароматический атом углерода, C.3 – sp^3 гибридный атом углерода, полное описание типов атомов см. в документации), номер и имя подструктуры, содержащей атом (если задана, на рис. 40 они присутствуют), заряд атома (если есть – его тип указан в @<TRIPOS>MOLECULE, на Рис. 40 в описании молекулы стоит NO_CHARGES и они не приведены), а также может содержать некоторую служебную информацию. Запись @<TRIPOS>BOND содержит расширенное описание таблицы связей молекулы и номер связи, номера двух атомов, между которыми образована связь, тип связи (1, 2, 3 – одинарная, двойная и тройная соответственно, ar-ароматическая, и другие), а также может содержать служебную информацию.

Детальное описание формата доступно на сайте компании Tripos Inc[35].

2.6.3. Формат базы данных белков PDB

Формат PDB был разработан для описания кристаллической структуры биологических макромолекул, таких как белки, нуклеиновые кислоты и их комплексы, а также связанных с ними экспериментальных данных. Формат развивался, и в настоящее время позволяет сохранять данные других экспериментальных методов, таких как ЯМР и криомикроскопия, а также теоретических расчётов. Информация по последней версии файла доступна на сайте базы данных белков PDB [86].

[35] http://tripos.com/data/support/mol2.pdf

Поскольку PDB-формат вошел в употребление еще в 1977 году [84] и первоначально был предназначен для ввода данных при помощи перфокарт, длина строки в нем не должна превышать 80 символов, в противном случае ставится знак /. В PDB-формате фиксировано положение полей внутри строки. Первые 6 символов каждой строки предназначены для ключевых слов, символы с 7 по 70 содержат данные, тогда как последние 10 символов обычно не используются. Ключевые слова в зависимости от типов данных могут быть либо ненумерованными (например, HEADER – название молекулы), либо нумерованными (например, ATOM_(6 пробелов)_1 – описание первого атома). Описание некоторых ключевых слов приведено в табл. 12.

Таблица 12

Секции и ключевые слова в PDB файле [84]

Секция	Описание	Запись в PDB
Заголовок	Общие описательные заметки (заголовки статьи, ключевые слова, авторы работы, журнал и т.п.)	HEADER, TITLE, COMPND, SOURCE, KEYWDS, AUTHOR, JRNL
Заметки	Библиография, заметки об уточнении структуры соединения	REMARK1, 2, 3 и т.д.
Первичная структура	Аминокислотная или нуклеотидная последовательность	SEQRES
Гетерогенная группа (гетероген)	Описание нестандарных групп (необычные аминокислоты, закомплексованные молекулы в активных сайтах биомолекул и др.)	HET, HETNAM, FORMUL
Вторичная структура	Описание вторичной структуры биомолекулы (какая часть биомолекулы имеет β-складчатую, α-спиральную структуру и т.п.)	HELIX, SHEET, TURN
Фрагментная связанность	Описание химической связанности внутри и между фрагментами (дисульфидные, водородные связи, межмолекулярная связь с гетерогеном	SSBOND, LINK, HYDBND
Особенности	Особенности внутри молекулы	SITE

Кристаллографические характеристики	Описание кристаллографической ячейки	CRYST1
Координатная трансформация	Операторы координатной трансформации	ORIGXn, SCALEn, MTRIXn, TVECT
Координаты	Данные о координатах атомов: стандартных (MODEL, ATOM) и нестандартных (HETATM) фрагментов	MODEL, ATOM, SIGATOM, HETATM
Атомная связанность	Указание ковалентно связанных атомов (без указания порядков связей)	CONECT
Разметка	Общая информация, указание окончания файла	MASTER, END

Поскольку формат разработан, прежде всего, для макромолекул, в файле может содержаться, помимо координат атомов, последовательность аминокислот, тип вторичной структуры, координаты молекул воды или других соединений, кристаллизующихся совместно с белками, а также дополнительная информация об условиях эксперимента, приборах, типе кристаллической упаковки, авторах, а также ссылки на публикации.

Спецификация биологической макромолекулы должна содержать описание последовательности аминокислот или нуклеиновых оснований в биополимере, в строке с ключевым словом SEQRES, после которой должен быть указан номер строки и число остатков в цепи. Остатки приводятся с N-конца для пептидов и с 5’-конца для нуклеиновых кислот с использованием общепринятого трехбуквенного обозначений для аминокислот и одно- или двухбуквенных обозначений нуклеиновых оснований (например, Т-тимин). Строка с описанием атома начинается с ключевого слова ATOM (для атомов, принадлежащих биополимерам) или HETATM (для малых молекул) и содержит: его номер, уникальный идентификатор атома внутри остатка (обычно представляет собой символ химического элемента, но для атомов углерода с указанием удаленности от карбонильного углерода – α-атом углерода как

CA, β-углерод – как CB и т.д.[36]), имя и номер остатка (аминокислоты или нуклеинового основания), которому принадлежит атом, его *x, y, z*-координаты в Å, заселенность (число эквивалентных атомов в данном кристаллографическом положении), фактор температурных колебаний, символ химического элемента и заряд (если есть).

2.6.4. Формат данных CML

Язык химической «разметки» CML [87-92] (от англ. *Chemical Markup Language*) является расширением более общего и универсального языка XML (от англ. *Extensible Markup Language*)[37]. CML был создан для обмена и управления химической информацией, содержащейся в базах данных. Этот достаточно универсальный язык может использоваться для представлений молекул, реакций, спектральной информации, аналитических и кристаллографических данных. Кроме того, этот формат позволяет задавать двухмерные и трехмерные координаты атомов.

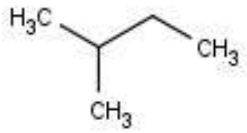
Формат позволяет хранить информацию о множестве молекул в одном файле. Представление CML организовано строго иерархически: молекула <molecule> → характеристика (например, атомы <*atomArray*>, связи <*bondArray*> и т.д.) → их описание (идентификационный номер *id*, вид атома *elementType*, стереохимия *parity*, заряд *formalCharge*, двумерные *X2, Y2* или трехмерные координаты *X3, Y3, Z3*). Конец указания данных обозначается косой чертой «/» с соответствующим индикатором. На рис. 41 приведено представление CML для молекулы 2-метилбутана.

CML является очень мощным инструментом для создания представления молекул и реакций, а также для организации связанных с ними данных. Он поддерживается большим количеством программ и интернет-ресурсов.

---

[36] Удаленность группировки по углеродной цепи от конкретной группы (чаще всего, кислотной) в органической химии иногда обозначается с использованием греческих символов. α-атомы углерода – это атомы, удаленные от данной (карбоксильной) группы на одну связь, β – на две, γ – на три и т.д. по греческому алфавиту.

[37]Подробнее про XML и другие языки можно прочитать на сайте Всемирного сетевого консорциума http://www.w3.org/




```
<?xml version="1.0"?>
<cml xmlns="http://www.xml-cml.org/schema"
xmlns:convention="http://www.xml-cml.org/convention" convention="convention:molecular"
xmlns:marvin="http://www.chemaxon.com/marvin/marvinDictRef">
<molecule id="m1">
 <atomArray>
  <atom id="a1" elementType="C"/>
  <atom id="a2" elementType="C"/>
  <atom id="a3" elementType="C"/>
  <atom id="a4" elementType="C"/>
  <atom id="a5" elementType="C"/>
 </atomArray>
 <bondArray>
  <bond atomRefs2="a1 a2" order="1"/>
  <bond atomRefs2="a1 a3" order="1"/>
  <bond atomRefs2="a1 a4" order="1"/>
  <bond atomRefs2="a2 a5" order="1"/>
 </bondArray>
</molecule>
</cml>
```

Рис. 41. Представление CML для 2-метилбутана

2.7. КОНВЕРТАЦИЯ МЕЖДУ ПРЕДСТАВЛЕНИЯМИ

Работа с разными программами, публикация, анализ и обработка данных требуют от химика умения их конвертировать из одного формата в другой. Например, химик-синтетик для публикации результатов экспериментов должен приводить для всех структур названия в соответствии с правилами ИЮПАК. В то же время для хранения данных о большом числе соединений и обмена структурами ему будет удобно использовать текстовые строки в формате SMILES или текстовые файлы в формате SDF. Если же требуется визуализировать молекулу и провести анализ ее реакционной способности, то ему нужно определить экспериментально либо рассчитать теоретически трехмерную структуру молекулы, которую он может хранить в MOL-формате.

Как уже отмечалось, структура молекулы может быть представлена на трех уровнях обобщения:

– 1D (элементный состав молекулы);

– 2D (связность атомов и стереохимическая конфигурация);

– 3D (конформация молекулы).

Переход на более высокий уровень обобщения, т.е. из 3D в 2D и из 2D в 1D, тривиален[38]. Действительно, не составляет труда, исходя из двумерной структуры молекулы, рассчитать ее элементный состав. Кроме того, зная трехмерную структуру молекулы, легко найти ее связанность. Более того, на основании определенных правил можно даже предположить формальные порядки связей с высокой степенью достоверности. В то же время конвертация представления низшего уровня в более высокий, т.е. из 1D в 2D и из 2D в 3D, представляет существенную проблему ввиду наличия, как правило, чрезвычайно большого количества вариантов. Проведение перебора таких вариантов принято называть *генерацией* (англ. *generation*) или *перечислением* (англ. *enumeration*). При конвертации из 1D в 2D количество вариантов (которое соответствует числу изомерных молекул) растет крайне быстро (сверхэкспоненциально) с ростом числа атомов и их типов. На практике это не позволяет сгенерировать все структурные изомеры для органических молекул, содержащих более 14 тяжелых атомов[39]. Не менее сложной проблемой является также конвертация из 2D- в 3D-представления, поскольку описываемой одной структурной формулой органической молекуле может соответствовать большой набор конформеров, а также бесконечно большое множество конформаций. Возможности конвертации между представлениями разного уровня обобщения схематично представлены на рис. 42.

Не все представления даже одного уровня обобщения могут быть конвертированы друг в друга. Для того чтобы можно было конвертировать представление, необходимо, чтобы из него могла быть восстановлена структура молекулы на данном уровне обобщения. По этой причине, например, невозможно из молекулярного отпечатка создать строку SMILES соединения. Более того, невозможно конвертировать

---

[38] Уровень 1D является более высоким уровнем обобщения, чем 2D, поскольку одному элементному составу (1D) может соответствовать несколько структурных формул (2D). Таким образом, одно описание элементного состава как бы «обобщает» множество структурных формул изомерных молекул. Аналогично, одна структурная формула (2D) «обобщает» множество конформеров (3D), и поэтому уровень 2D является более высоким уровнем обобщения, чем 3D.

[39] В теоретической химии тяжелыми атомами называют все атомы кроме водорода и гелия.

молекулярные отпечатки разных типов друг в друга. К сожалению, даже в тех случаях, когда конвертация представлений возможна, она часто бывает осложнена множеством проблем.

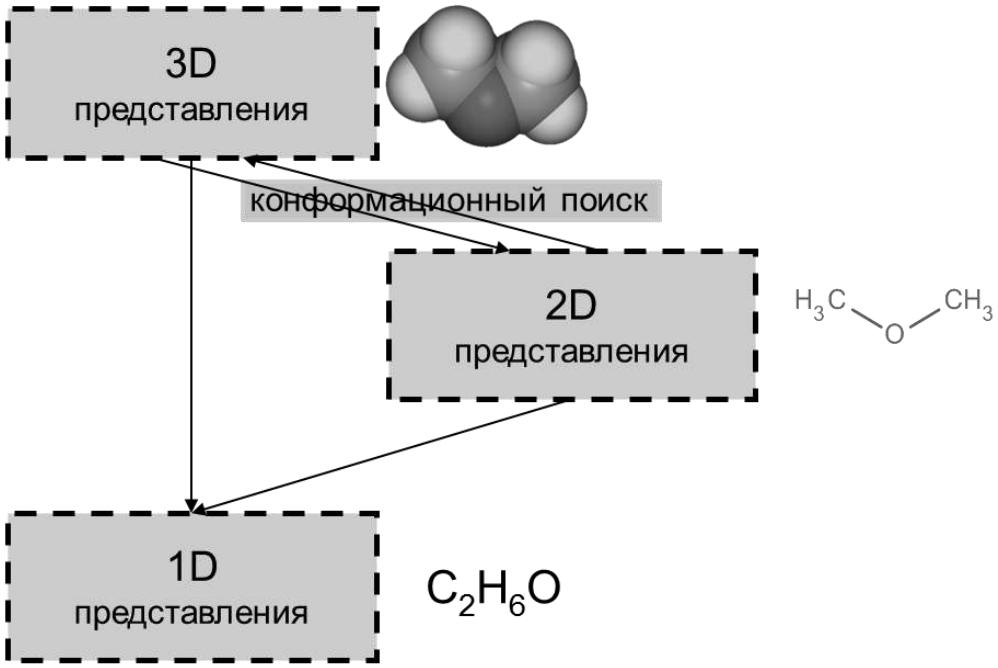


Рис. 42. Возможности конвертации между представлениями различного уровня

Далее мы рассмотрим наиболее важные виды конвертации между представлениями, возникающие при этом проблемы, их решения, а также необходимые для этого программы.

2.7.1. Конвертация структура-линейное представление

Существует множество программ, позволяющих конвертировать структуру в разнообразные линейные представления. Распространенные виды линейных представлений (SMILES, SLN, InChI) оптимизированы для компьютерной обработки. Кроме того, они учитывают структурные особенности, типично встречающиеся в органических, неорганических и элементоорганических соединениях. В то же время конкретные программы-конвертеры могут обладать определенными ограничениями. Конвертеры от компаний-разработчиков (Daylight, Tripos и IUPAC, соответственно) максимально работоспособны и поддерживают все задокументированные особенности линейных представлений. В то же время их некоммерческие аналоги (за исключением

InChI), либо коммерческие реализации, разработанные в других компаниях, могут обладать неполной функциональностью.

2.7.1.1. Конвертация структура-систематическое имя

При работе с химической информацией чрезвычайно важно уметь составлять для молекулы систематическое имя. Правила ИЮПАК [93], несмотря на их сложность, рекомендованы к использованию в химической литературе и других источниках химической информации, поскольку название соединения в систематической номенклатуре несмотря на сложность понятно любому химику. Другим важным аспектом является то, что названия, составленные по правилам ИЮПАК, используются во многочисленных базах данных, патентах и документах для однозначного описания химических соединений в понятной человеку форме. Введение нового соединения в базу данных часто требует создания его систематического имени. Поэтому важным является наличие автоматизированной системы, которая бы генерировала систематическое название в номенклатуре ИЮПАК и, наоборот, преобразовывало название в структуру.

Следует подчеркнуть, что составление систематического названия для химического соединения является крайне сложной задачей. Причин для этого достаточно много. Во-первых, в некоторых случаях правила ИЮПАК допускают существование нескольких альтернативных имен для соединений, выбор между которыми зачастую определяется предпочтением научного работника. Во-вторых, правила ИЮПАК сложны и не всегда однозначны, и потому не могут быть в полном объеме алгоритмизированы. В-третьих, существуют разные наборы правил номенклатуры для различных классов соединений. В-четвертых, для лучшего понимания химиками используется большое число заместителей и основных структурных фрагментов с тривиальными (несистематическими) названиями. Все это привело к тому, что производители программного обеспечения для основных систем индексирования данных, в частности CAS и Beilstein Institute, разработали свои варианты алгоритмов для составления систематических имен соединений, основанные на несколько пересмотренных и, к сожалению, не задокументированных правилах систематической номенклатуры. Вследствие этого генерируемые ими систематические названия для одного и того же соединения могут отличаться. Поэтому в Отделении ИЮПАК по химической номенклатуре и представлению структур был

инициирован проект, имеющий целью убрать неоднозначности номенклатуры и создать правила по генерации только одного уникального систематического названия соединения. Результатом проекта стало издание в 2005 году «Предварительных рекомендаций ИЮПАК» [94][40], согласно которым названия, созданные в соответствии с этими рекомендациями, называются *предпочтительными* (англ. *preferred names*), а все другие названия в рамках номенклатуры ИЮПАК – *общими* (англ. *general names*).

Конвертация систематического названия в структуру

Конвертация систематического названия в структуру и структуры в название происходит с разной степенью сложности, и при этом используются различные алгоритмические подходы. Конвертация названия в структуру – задача более легкая, хотя и не совсем тривиальная, поскольку название может быть дано на любом из «диалектов» номенклатуры ИЮПАК, а также с некорректным синтаксисом. Для конвертации названия в структуру создается библиотека фрагментов, каждому из которых соответствуют определенные сочетания букв в имени соединения, т.н. морфемы (например, CH_3 – «метил», «мет»). Общий принцип конвертации систематического названия в структуру заключается в его разбиении на текстовые фразы максимальной длины, которые подвергаются лексическому анализу и выявлению морфем. Далее, исходя из морфем, находятся подструктурные фрагменты, соответствующие как «скелету» химической структуры, так и заместителям, которые затем состыковываются в соответствии с их расположением, указанном в названии.

Конвертация структуры в систематическое имя

Алгоритм конвертации структуры в систематическое название чрезвычайно сложен, поскольку должен работать в соответствии с полным набором правил номенклатуры ИЮПАК, причем из множества возможных названий должно быть выбрано только одно. В 2002 году официально[41] было признано, что дизайн, решения и ме-

---

[40] Доступен по адресу: http://www.iupac.org/fileadmin/user_upload/publications/recommendations/CompleteDraft.pdf

[41] На Конференции CAS/IUPAC по химическим идентификаторам и использованию XML в химии, Коламбус, Огайо, США, июль 2002 года.
http://old.iupac.org/symposia/conferences/CIandXML_jul02/index.html

тоды алгоритма *AutoNom* [95-99] (разрабатывался для Beilstein Institute) являются стандартными для преобразования структуры молекулы в систематическое название. На первом этапе работы алгоритма производится поиск наименьших наборов наименьших циклов (англ. *Smallest Sets of Smallest Rings* – SSSR)[42], которые потом объединяются в более крупные блоки, которым присваиваются систематические названия. Далее производится поиск функциональных групп и их ранжирование в соответствии с правилами старшинства. После этого выбирается самая старшая группа, которая войдет в название родительского соединения. На основании всей полученной информации выбирается родительская основа и заместители, и атомы в основе нумеруются в соответствии с правилами ИЮПАК. Знание родительской основы и заместителей позволяет построить так называемое *дерево названия* – граф, корнем которого является родительская структура, ветви – заместителями, которым еще не присвоены систематические названия по правилам ИЮПАК, листья – морфемами. Далее, продвигаясь от корня дерева названия к листьям, составляются из морфем и заносятся в таблицу названия фрагментов. Затем, продвигаясь в обратном направлении, названия фрагментов читаются из таблицы и объединяются в более крупные фрагменты до тех пор, пока не будет достигнут корень дерева. После этого проводятся специальные процедуры улучшения (упорядочение по алфавиту, расстановка пунктуации), и на основании полученного дерева строится название соединения.

Программы взаимной конвертации систематического имени и структуры

К настоящему времени разработано несколько программ, выполняющих преобразование структуры в систематическое название и обратно.

---

[42] Подчеркнем, что алгоритм поиска наименьшего набора наименьших циклов в графах уже сам по себе довольно сложный, и на разработку его вариантов были затрачены большие усилия математиков и программистов.

- API-модуль *molconvert* и программа Marvin Sketch от ChemAxon [100] (поддерживают также предпочтительные систематические имена[43]);
- модуль *Struct=Name Pro* в составе программного комплекса ChemOffice от компании CambridgeSoft (вошла в состав Perkin Elmer) [101];
- пакет *ACD/Name* и ее batch-модули от компании ACD Labs [102];
- NamExpert и Nomenclator от компании ChemInnovation [103];
- Symyx Draw от Accelrys [104];
- ограниченные возможности по конвертации предоставляет программа DrawIt от Bio-Rad Laboratories [105].

Тем не менее еще рано говорить о существовании универсальной и идеальной программы, однозначно выдающей имя в полном соответствии с номенклатурой ИЮПАК. Авторы большинства программ вынуждены для упрощения алгоритмизации вводить определенные ограничения или даже отступления от правил ИЮПАК. Таким образом, проблема взаимной конвертации структуры химического соединения и его систематического названия в настоящее время полностью еще не решена.

2.7.2. Конвертация двухмерной структуры в трехмерную

Многие биологические, физические и химические свойства молекул явным образом зависят от трехмерной структуры молекулы. Рассмотрение лиганд-рецепторных взаимодействий, некоторых молекулярных свойств, химической активности требуют не только информации о том, как атомы связаны, но и о том, как они расположены друг относительно друга в трехмерном пространстве. Вместе с тем лишь несколько экспериментальных методов позволяют определять трехмерную структуру молекул: рентгеноструктурный анализ (РСА), ядерный магнитный резонанс (ЯМР), микроволновая спектроскопия, электронная и нейтронная дифракция в газовой фазе. Крупнейшая база трехмерных структур молекул – Кембриджская структурная база дан-

[43] Обращаем ваше внимание, что авторы программы не гарантируют 100% соответствия рекомендациям ИЮПАК 2004 года. http://www.chemaxon.com/marvin/help/calculations/s2n.html

ных (CSD) [106-109] – содержит чуть более 1 072 000 структур[44], что составляет менее 1% от числа известных веществ[45]. Кроме того, трехмерные структуры числом около 174 000 биологических макромолекул (главным образом белков) содержатся в Банке данных белков (PDB) [110]. Таким образом, для абсолютного большинства химических структур трехмерная структура неизвестна. С другой стороны, для проведения виртуального скрининга в большинстве случаев требуется знание трехмерной структуры молекул. Вследствие этого конвертация 2D-представлений в трехмерные является ключевой задачей хемоинформатики. Применяемые для этого программы должны удовлетворять множеству критериев: устойчивости работы, высокой скорости генерации структуры, учета стереохимии, высокого качества результатов, возможности работы с широким спектром соединений.

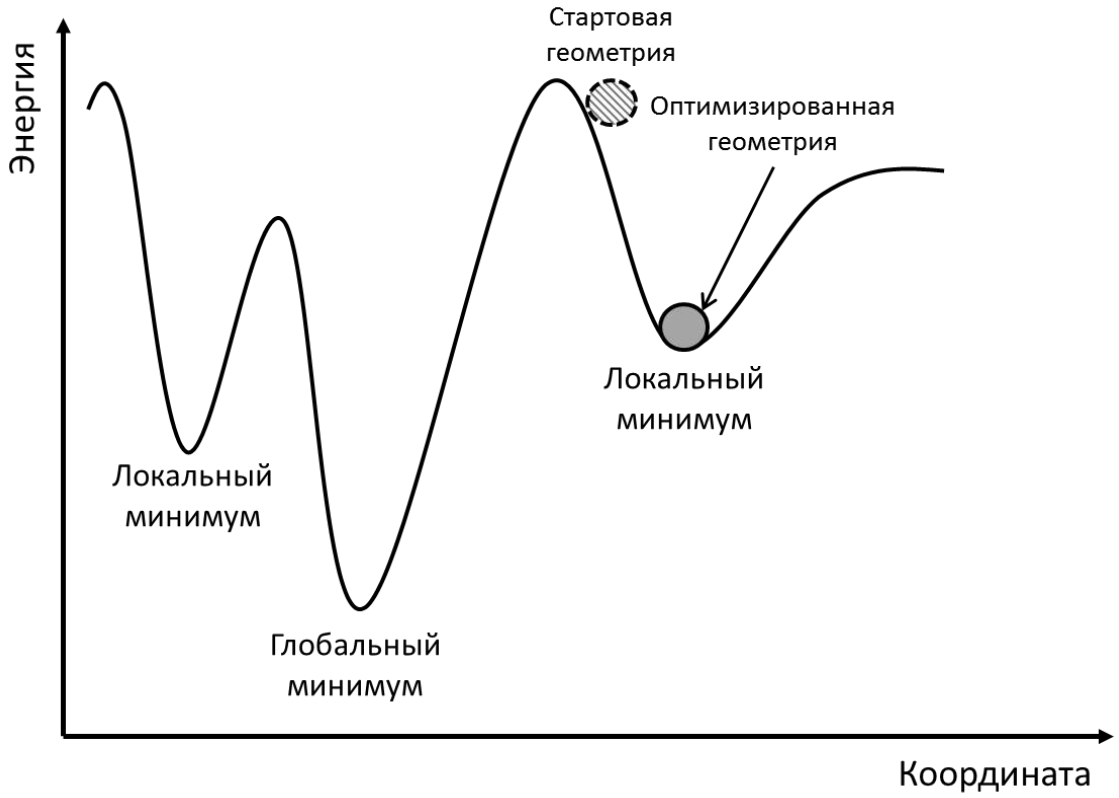


Рис. 43. Локальный и глобальный минимумы на поверхности потенциальной энергии молекулы. Локальная оптимизация геометрии приводит к структуре ближайшего локального минимума

[44] По данным на 06.07.2020.

[45] Согласно базе CAS известно 162 000 000 соединений по данным на 06.07.2020

Как уже отмечалось, существует бесчисленное число вариантов реализации трехмерной структуры молекулы, т.е. конформаций. Не все из них, однако, могут реально существовать. Дело в том, что в соответствии с адиабатическим приближением Борна-Оппенгеймера каждой пространственной конфигурации молекулы (т.е. каждому набору внутренних координат – длин связей, валентных и торсионных углов) можно приписать конкретное значение энергии – т.н. адиабатический потенциал. При изменении пространственной конфигурации молекулы значение этого потенциала меняется, образуя в пространстве внутренних координат системы так называемую поверхность потенциальной энергии. В этом случае происходящие при конформационных переходах изменения трехмерной геометрии молекулы могут быть описаны как движение «шарика» по поверхности (гиперповерхности) потенциальной энергии[46]. Точка на этой поверхности, соответствующая самому минимальному значению энергии, называется *глобальным минимумом*. У находящейся в нем молекулы любые изменения внутренних координат приводят только к повышению энергии (рис. 43).

Для большинства молекул на поверхности потенциальной энергии существуют также менее глубокие минимумы, называемые *локальными*. У находящихся в них молекул, т.н. конформеров, лишь небольшие изменения всех внутренних координат приводят к повышению энергии, и есть шанс при определенном изменении координат понизить энергию. Таким образом, глобальный минимум является самым глубоким (т.е. характеризующимся самым низким значением энергии) из локальных минимумов на поверхности потенциальной энергии молекулы, а находящейся в нем конформер – наиболее устойчивым. Поиск локального и глобального минимума – чрезвычайно важная задача для молекулярного моделирования, поскольку молекула проводит в состоянии, близком к глобальному минимуму, большую часть времени, а следовательно, данное состояние наиболее хорошо описывает молекулу, т.е. оно наиболее репрезентативно. Таким образом, задача конвер-

---

[46] Поверхность в n-мерном пространстве внутренних координат молекулы (X_1, $X_2 \ldots X_n$), каждая точка которой по оси ординат (оси Y) равна потенциальной энергии молекулы, называется поверхностью потенциальной энергии. Иногда используют слово гиперповерхность, чтобы показать, что размерность пространства внутренних координат больше 3.

тации 2D-структуры молекулы в трехмерную сводится к поиску ее наиболее устойчивого конформера (рис. 43).

В ряде случаев возникает необходимость знать не один наиболее стабильный конформер для молекулы, а конформации молекул в локальных, или не сильно от них отличающихся, минимумах. Это позволит предсказывать, например, расположение молекулы лекарства в активном центре фермента, или судить о динамике молекул. Для этих целей следует рассматривать не только глобальный, но и локальные минимумы на поверхности потенциальной энергии молекулы и даже конформации, удаленные от минимума. Эти конформации, как правило, существенно отличаются друг от друга только значениями диэдральных (торсионных) углов. Вследствие этого для перебора конформаций одной молекулы обычно ограничиваются поиском по изменяющимся диэдральным углам. В соответствии с распределением Больцмана, для молекулы вероятность принимать заданную конформацию падает экспоненциально с ростом ее энергии[47]. Вероятность найти конформацию, отличающуюся по энергии от самой стабильной на 1 ккал/моль (4.18 кДж/моль), при 298 К в 5.5 раз меньше, чем наиболее стабильный. Для конформаций, отличающихся по энергии на 10 ккал/моль, присутствует 1 молекула менее стабильной конформации на 20 млн более стабильной. Таким образом, при обычных условиях вероятность найти подобные малостабильные конформации является очень низкой, и их присутствие никак не влияет на свойства системы. Вследствие этого поиск конформаций обычно ограничивают определенным *энергетическим интервалом* (англ. *energy gap*). Конформации, чьи относительные энергии попадают в этот интервал, считаются реализуемыми, другие отбрасываются как маловероятные. Таким образом, задачей конформационного поиска является получение набора определенного числа (часто задаваемого пользователем) наиболее стабильных и при этом достаточно сильно отличающихся друг от друга конформаций.

Актуальной проблемой для хемоинформатики является задача поиска так называемой биоактивной конформации. Дело в том, что в

---

[47] Закон распределения Больцмана для числа частиц N_1 и N_2 имеющих энергии E_1 и E_2 (выраженных в Дж/моль) записывается так:

$$\frac{N_1}{N_2} = \exp\left(\frac{E_2 - E_1}{RT}\right)$$

комплексе с биологической макромолекулой (например, внутри активного центра фермента) конформация молекулы может сильно отличаться от стабильных конформеров, найденных для этой молекулы в свободном состоянии (т.е. в вакууме либо в водном растворе). Причина этого заключается в том, что в данном случае вероятность найти молекулу в заданной конформации определяется уже не минимумом ее потенциальной энергии, а минимумом свободной энергии ее комплекса с макромолекулой. Поскольку последняя величина крайне сложно поддается даже приближенной оценке, для поиска «биоактивной» конформации необходимо генерировать большой набор разнообразных конформаций для весьма широкого энергетического интервала, причем может быть совершенно не важно, соответствуют ли они минимумам на поверхности потенциальной энергии для молекулы.

Можно очень условно выделить несколько подходов для автоматической конвертации двухмерной структуры в трехмерную:

- методы, основанные на правилах и данных,
- фрагментные методы,
- методы конформационного поиска,
- методы молекулярного моделирования,
- метод следования минимальной моде колебаний,
- методы метрической геометрии.

Каждый из упомянутых подходов обладает своими преимуществами и недостатками. Иногда достаточно сложно отнести программу строго к одному методу, поскольку они (программы), как правило, используют одновременно несколько подходов либо идеологически могут быть по-разному интерпретированы. Например, методы конформационного поиска ставят своей целью генерацию множества конформаций, из которой потом выбирается одна наилучшая. По сути, все указанные методы включают элементы конформационного поиска. Каждая программа генерации трехмерной структуры так или иначе должна выбирать одну структуру из набора возможных, поскольку в общем случае невозможно определить глобальный минимум, не перебрав множество локальных.

Все приведенные методы взаимосвязаны, и, в конечном итоге, все они используются для генерации набора конформеров (конформационного поиска), поэтому мы не будем проводить искусственной границы между генерацией трехмерных структур и конформационным поиском, имея в виду, что, по сути, задача отличается только количеством выдава-

емых конформеров. Подробнее с описанием методов генерации трехмерных структур можно ознакомиться в главе И. Садовски [111], а про методы конформационного поиска – в главе К. Шваба [112].

2.7.2.1. Методы, основанные на правилах и данных

Методы, основанные на правилах и данных, базируются на эмпирических знаниях химиков относительно того, как должна выглядеть трехмерная структура молекулы. Основная идея этих методов заключается в конструировании молекулы из конформационных элементов с хорошо известными оптимальными геометриями (например, кресло для циклогексана), которые далее состыковываются с использованием эмпирических правил.

Для этого в программах *Wizard* [113-117] и *Cobra* [118, 119] молекула разбивается на перекрывающиеся подструктурные фрагменты (циклы и короткие цепочки), каждому из которых на основании структур шаблонов, хранящихся в памяти, приписывается та или иная конформация. Далее набор конформеров комбинируется в виде графа-дерева, в каждом из узлов которого находится одна из возможных конформаций фрагмента. Отбор предпочтительных конформеров проводится при помощи алгоритма A* [120] с использованием специальной функции «потерь», которая в числовом виде выражает предпочтительность/нежелательность данной конформации фрагмента на основании правил, выработанных химиками, и результатов машинного обучения. В данном случае используемая функция имеет смысл энергии, подлежащей минимизации. Далее фрагменты по специальным правилам накладываются на молекулу, что приводит к получению трехмерных координат для входящих в нее атомов. Из сгенерированных таким образом конформаций отбирается наилучшая по критерию минимизации значений функции «потерь». В алгоритме A* предусмотрен также механизм, приводящий к генерации максимально различных конформеров. Подобные методы в вычислительном плане чрезвычайно эффективны, однако качество получаемых конформаций сильно зависит от того, содержатся ли в базе данных все необходимые для работы с данной молекулой шаблоны.

Программа *Concord*[48] [121, 122] разбивает молекулу на фрагменты (циклы и ациклические атомы), и длинам связей и валентным углам

[48] Является частью программы Sybyl-X компании Tripos.

присваиваются табличные значения в зависимости от типа атомов. Циклам присваиваются возможные конформации, из которых ищется конформация с минимальным значением функции напряжения. При этом используется специальная процедура оптимизации системы колец одно за другим. Конформация ациклической части находится путем минимизации функции энергии, включающей рассмотрение всех 1-4, 1-5 и 1-6 взаимодействий и контактов несвязанных атомов. Таким образом, этот метод содержит в себе элементы методов силовых полей, и поэтому подходит в основном для органических молекул.

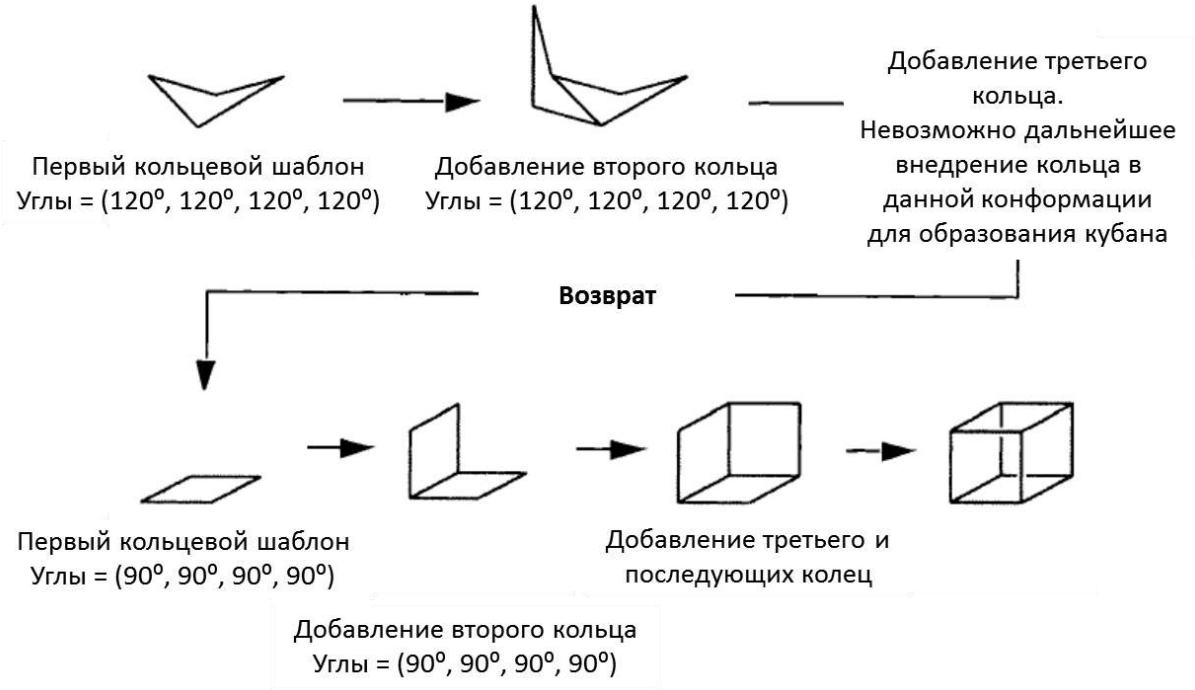


Рис. 44. Использование алгоритма возвратов в программе Corina для построения конформации кубана

Программа *Corina* [123-126][49] применима к разнообразным классам органических соединений, включая макроциклы и металлорганические соединения. На первом этапе генерации трехмерной структуры длинам связей и валентным углам присваиваются табличные значения. Далее в сопряженных системах делаются поправки в соответствии с методом молекулярных орбиталей Хюккеля[50]. После этого молекула

[49] Доступна по адресу: http://www.molecular-networks.com-/products/corina, on-line версия: http://www.molecular-networks.com-/online_demos/corina_demo

[50] Подробнее о методе Хюккеля можно прочесть в учебниках по квантовой химии, например, Гл. 7.5.2 и Гл. 8 в *Минкин В.И., Симкин*

разбивается на ациклические атомы и кольца, причем к последним относят также атомы, которые непосредственно связаны с кольцами (это позволяет учесть их влияние на конформацию), причем конденсированные системы считаются одним кольцом. С использованием *алгоритма поиска с возвратом* (англ. *backtracking algorithm*) определяется конформация конденсированных колец путем такого комбинирования конформаций малых колец-шаблонов, чтобы суммарная энергия напряжения, рассчитываемая из энергий напряжений шаблонов, была минимальной (рис. 44).

Конформация полимакроциклов рассчитывается на основании принципа суперструктуры – разбивая цикл на более мелкие фрагменты, рассчитывая их конформацию и объединяя в единую структуру. Возможные конформации больших колец строятся с использованием символического представления циклов Дэйла [127], и полученные конформации далее ранжируются по энергии. Оптимизация геометрии циклов проводится при помощи псевдо-силового поля, переходя последовательно от более жестких к менее жестким кольцевым структурам в молекуле. Конформации ациклических фрагментов и их «энергии напряжения»[51] определяются в соответствии с более чем 900 правилами, хранящимися в так называемой библиотеке углов и основанными на статистическом анализе рентгеноструктурных данных. Конформации ациклических и циклических фрагментов комбинируются, и выбирается трехмерная структура с наилучшим значением псевдоэнергии. При наличии дефектов в пространственной структуре молекулы на последнем шаге может проводиться дополнительная оптимизация геометрии при помощи силового поля, параметризованного с использованием рентгеноструктурных данных.

Программа Corina может работать практически со всеми химическими элементами из таблицы Менделеева, что позволяет с ее помощью генерировать пространственную структуру металлоорганических соединений. В вычислительном плане она очень эффективная и позво-

---

Б.Я., Миняев Р.М. Теория строения молекул. Ростов на/Д: Феникс, 1997. 560 с.

[51] Это не есть энергия в чистом виде, а некий штраф за стерическое напряжение, которое имеет смысл энергии стерического напряжения.

ляет работать с достаточно крупными молекулами с 700 атомами. Тем не менее программа не принимает во внимание существование внутримолекулярных взаимодействий и водородных связей. Для расчета конформаций (в особенности биоактивных) на основании данных программы Corina используется программа Mimumba [128, 129]. Для того, чтобы генерировать биоактивные конформации, библиотека углов расширяется с использованием структур трипептидов из PDB. Для генерации конформеров на этапе комбинирования ациклических и циклических фрагментов отбрасываются те структуры, которые имеют слишком большую (определяемую пользователем) «энергию» или совпадают по «энергии» в пределах указанного пользователем допуска. Программа *Rotate*[52] [130], также созданная для конформационного поиска, использует «библиотеку углов» Corina и полученную ею конформацию циклов. Для генерации конформаций Rotate отбирает значения диэдральных углов, лежащие ниже определенного значения, которые далее используются для генерации пространственной структуры. Для ограничения количества конформаций они кластеризуются на основании среднеквадратичного отклонения их декартовых координат или диэдральных углов. Пороговое значение, на основании которого молекулы относятся к одному классу, определяется пользователем.

Программа *OMEGA* [131] – очень популярный в настоящее время способ генерации конформеров. Она, так же как и вышеуказанные программы, разбивает молекулу на отдельные экзоциклические и циклические фрагменты. Конформация этих фрагментов находится на основе предопределенной библиотеки фрагментов. Далее они объединяются таким образом, чтобы по возможности учесть все конформации циклических фрагментов. В молекулах определяются все свободно вращающиеся (одинарные) экзоциклические связи, которым присваиваются, на основании заложенных в программу правил, возможные значения диэдральных углов, и далее проводится систематический поиск по всем их возможным комбинациям. Окончательно конформеры выбираются с учетом или заданного пользователем их максимального количества или величины энергетического окна. В последнем случае применяется модифицированное силовое поле Дрейдинга, причем отбираются только конформеры, среднеквадратичное отличие между ко-

[52] Доступна по адресу: http://www.molecular-networks.com-/products/rotate

ординатами атомов которых превышает заданную пользователем величину. Программа работает быстро, позволяет генерировать конформеры для большого числа соединений одновременно. Кроме того показано [132], что OMEGA хорошо воспроизводит биоактивные конформации, для генерации которых она широко и используется. Программа OMEGA неоднократно признавалась лучшей по соотношению качества результатов и затрат времени [112, 133].

Основанный на правилах подход реализован также в автономном модуле *Molconvert* и подключаемом вычислительном модуле (API-модуле) к программе Marvin Sketch[53] от компании ChemAxon [100]. Генерация трехмерной геометрии и поиск конформеров основан на сборке молекулы из фрагментов. Каждый фрагмент представляет собой полноценную трехмерную молекулу, в которой разорванные связи замещены водородами. Фрагменты пошагово укрупняются добавлением атомов к сгенерированным на более ранних шагах фрагментам до тех пор, пока не будет собрана искомая молекула. На каждом шаге рассматриваются несколько возможных конформаций фрагментов. Энергии конформеров рассчитываются с использованием модифицированного силового поля Дрейдинга [134]. Сборка фрагментов в искомую структуру производится с использованием определенного набора правил. Для сложных случаев, когда конформация фрагмента не может быть легко определена, используются подходы метрической геометрии (см. далее). Полученные структуры оптимизируются либо с помощью «классического» силового поля (Дрейдинга) [134], либо силовым полем MMFF94 [135-139]. Следует также иметь в виду, что отбрасывание «малозначимых» конформаций фрагментов может привести к тому, что структура, полученная даже при «точной» генерации трехмерной структуры (Clean3D>Fine), может оказаться не самой стабильной из найденных в конформационном поиске с использованием этой программы.

2.7.2.2. Методы, основанные на фрагментах

Идея этих методов состоит в том, чтобы строить трехмерную структуру исходя из структуры аналогов данной молекулы или состав-

---

[53] Описание метода доступно по ссылке: http://www.chemaxon.com/conf/Advanced_automatic_generation_of_3D_molecular_structures.pdf

ляющих ее фрагментов, трехмерные структуры которых известны и хранятся в базе данных.

Основная идея программы *AIMB* [140-143] заключается в том, что в крупных и достаточно разнообразных базах данных, содержащих полученные из эксперимента трехмерные структуры молекул, в неявном виде содержатся знания о том, как строить трехмерные модели других молекул. Так называемая база знаний AIMB содержит пространственные структуры множества малых (менее 65 тяжелых атомов) органических молекул. При построении трехмерной модели молекула разбивается на циклические и ациклические фрагменты таким образом, чтобы внутри фрагментов содержалось максимальное количество связей и минимальное – между ними. Если аналогов фрагментов нет в «базе знаний», молекула разбивается дальше, однако полициклические системы расщепляются только до самых малых циклов. Далее производится поиск аналогов данных фрагментов в «базе знаний» с заданным порогом на уровень сходства, который может уменьшаться до тех пор, пока не будет найдено 5-10 аналогов. С учетом разной приоритетности атомов и разной степени приближения (например, полное совпадение, несовпадение символов атомов, удаленных от места разрыва более чем на 2 связи, точно на 2 связи, или менее чем на 2 связи), а также учета валентности, гибридизации, типов атомов и других характеристик находится лучший аналог. Далее фрагменты, которым приписана трехмерная структура, объединяются в трехмерную модель молекулы. Этот метод является очень эффективным и позволяет быстро генерировать трехмерные структуры, близкие к глобальному минимуму на поверхности потенциальной энергии. Поскольку стереохимическая конфигурация фрагментов, найденных в базе знаний, может быть разная, этот метод позволяет также строить наборы различных стереоизомеров. При хорошем выборе базы данных метод неявно может принимать во внимание дальнодействующие и невалентные взаимодействия.

Близкий подход реализован также в программе *Chem-X* [144], разработанной в компании Chemical Design. Используемая в ней база данных содержит относительно небольшое число фрагментов, в том числе фрагменты с обобщенными типами атомов, с различной степенью ненасыщенности связей и стереохимической конфигурацией. Поисковая машина данной программы сначала ищет в библиотеке фрагменты, полностью совпадающие с имеющимися в молекуле, а если точного

совпадения не найдено, то используются хранимые в библиотеке обобщенные структуры. Полициклические структуры могут рассматриваться либо целиком, либо делиться на составляющие их циклы. Пространственная структура ациклических частей строится на основании табличных значений диэдральных углов. Программа работает очень быстро, однако дальнейшее ее усовершенствование за счет укрупнения библиотеки существенно ее замедляет. При использовании программы следует также иметь в виду, что слишком простой способ генерации боковых цепей может плохо учитывать возможное влияние дальнодействующих взаимодействий.

2.7.2.3. Методы конформационного поиска

Основной задачей конформационного поиска является перебор по возможности всех возможных конформаций для данной молекулы и выбор из них наилучшей в соответствии с заданными пользователем критериями. Примерами наиболее часто используемых в хемоинформатике критериев являются попадание в заданное энергетическое окно (т.е. энергия конформаций не должна превышать энергию глобального минимума на значение, превышающее определенный порог) и требование максимального разнообразия отбираемых конформаций.

Систематический поиск

Систематический поиск является самым простым для понимания, но в то же время самым тщательным типом поиска. К сожалению, для молекул, содержащих большое число свободно вращающихся одинарных связей, он не может быть применим в силу его крайней затратоемкости. Для осуществления поиска диэдральный угол для каждой способной к вращению одинарной связи систематически варьируется с шагом 360˚/n (n – определяемое пользователем число, чаще всего 6), при этом длины связей и валентные углы остаются постоянными. Метод хорошо работает на ациклических системах, однако не подходит для циклов ввиду наличия зависимости между диэдральными углами в них. Для перебора всех конформаций обычно строится так называемое поисковое дерево (рис. 45), узлы которого представляют собой возможные значения диэдрального угла для каждой способной к вращению связи. При работе программы сначала осуществляется движение по дереву от корня через самую младшую ветвь, пока оно не достигнет концевого узла, который представляет собой первую из сгенерированных конформаций. На рис. 45:

- корень дерева расположен сверху, поэтому движение происходит сверху вниз;
- самая младшая ветвь расположена слева, причем старшинство узлов определяется по значению соответствующих диэдральных узлов;
- концевой узел – самый младший из листьев дерева – расположен внизу.

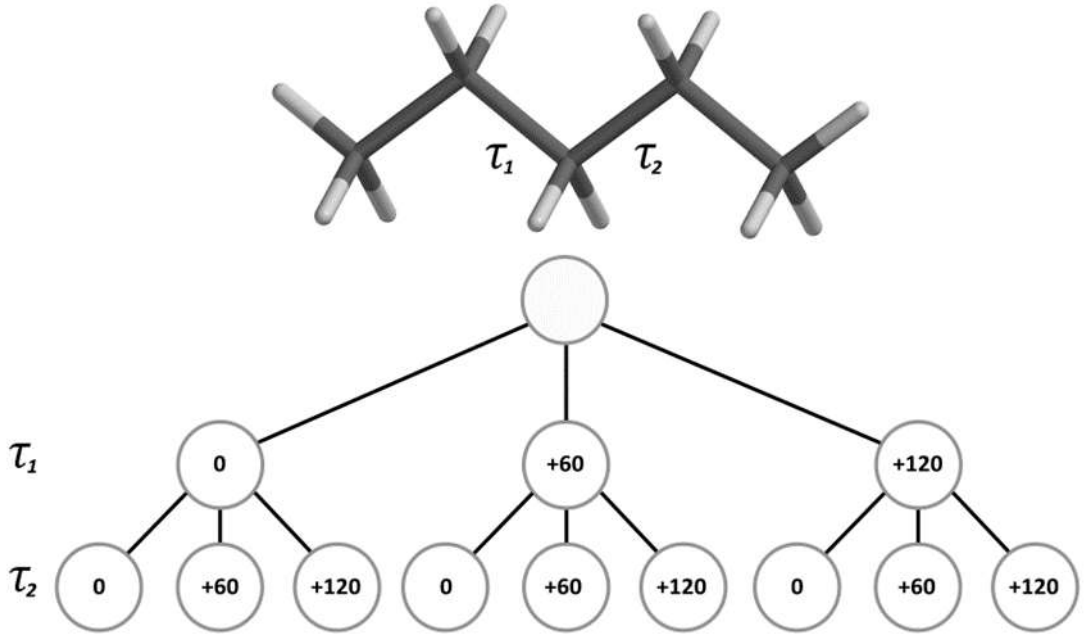


Рис. 45. Пример поискового дерева
в методах систематического конформационного поиска

Далее осуществляется возврат до уже пройденного промежуточного узла, из которого осуществляется движение по направлению к листьям через более старшую ветку. Это повторяется до тех пор, пока не будут перебраны в порядке старшинства все ветки, исходящие из данного промежуточного узла. После этого производится возврат к наиболее близко расположенному к корню узлу, для которого последовательно в порядке старшинства перебираются все исходящие из него ветки. Все это повторяется до тех пор, пока все дерево не будет пройдено. Этот алгоритм называется *поиском с возвратом*.

С увеличением числа вращаемых связей время работы программы резко возрастает (происходит т.н. *комбинаторный взрыв*), поскольку в общем случае должно быть перебрано n^k конформаций, где n – количество инкрементов (возможных значений угла), k – число вращаемых связей. Очевидно, что для эффективного использования систематиче-

ского поиска нужно максимально ускорить процесс. Наиболее простым способом является способ «*обрезки дерева*» (англ. *tree-pruning*). Для этого при достижении узла, перед тем, как спускаться ниже и начать сложную вычислительную процедуру оптимизации геометрии, проверяется, не возникают ли стерические затруднения в результате соответствующих данному узлу изменений в структуре, например, сильно сближенных атомов, невыгодных гош-взаимодействий. Если в результате обнаруживается такая проблема, то поиск возвращается на предыдущий узел. Используя такие правила, можно на ранних этапах (не достигая узлов-листьев поискового дерева) «обрезать» ветки и сэкономить существенное количество машинного времени.

Некоторые пакеты: Cerius^2, Vega ZZ, Spartan – включают возможности систематического конформационного поиска. Систематический перебор конформаций циклических структур представляет собой более сложную задачу, для решения которой предложено несколько подходов:

– на основе «псевдо-ациклического» рассмотрения молекулы (программа MULTIC [145]);

– путем отражения уголковых частей в плоскости ближайших атомов (т.н. «хлопанья углами» – англ. *corner flapping*), реализованного в программе CONFLEX [146].

Помимо программы OMEGA, упомянутой выше, процедура систематического поиска используется в одном из алгоритмов (т.н. *fast algorythm*) в программе ConFirm [147, 148] пакета Catalyst [149, 150] и Cerius^2 [151] (обе программы – продукты компании Accelrys[54]). Количество инкрементов в нем определяется гибридизацией связанных атомов, а каждая конформация проверяется на отсутствие «перекрываний» атомов, что в ряде случаев позволяет существенно сократить поиск.

Стохастический поиск

Как следствие происходящего с увеличением числа вращаемых связей и уменьшения инкремента вращения комбинаторного взрыва, рассмотренный выше систематический поиск реально осуществим только при очень небольшом (обычно берется не больше 3) числе вращаемых связей. В то же время значительная часть молекул содержит существенно большее число способных к вращению связей, что не позволяет для них проводить систематический перебор конформаций. В этом случае единственным работающим подходом является генера-

[54] http://www.accelrys.com

ция конформаций путем задания случайных значений варьируемым диэдральным углам, что эквивалентно «бросанию» точки в случайное место на поверхности потенциальной энергии молекулы. Если организовать разбрасывание точек таким образом, чтобы вероятность их попадания в данную область конформационного пространства зависела от близости к глобальному минимуму энергий соответствующих конформаций, то подобная процедура может быть весьма эффективна для нахождения наиболее стабильного конформера. Такой вид поиска обычно называют *стохастическим*. Методов стохастического поиска достаточно много, и применяются они не только в конформационном поиске, но и при решении других, самых разнообразных оптимизационных задач. К таким методам относятся метод Монте-Карло, генетический алгоритм [152], алгоритм муравьиной колонии (англ. *ant colony optimization algorithm*) [153-155], алгоритм улья (англ. *artificial bee colony optimization algorithm*) [156] и другие.

Генетический алгоритм является одним из методов, основанным на подражании биологическим процессам, в данном случае – процессу эволюции. Дадим краткую характеристику метода в приложении к задаче конформационного поиска. Изменяемые характеристики конформаций данной молекулы (в частности, диэдральные углы) объединяются в один набор чисел, который отождествляется с *хромосомой*, представляющей один генотип и, следовательно, одну особь в популяции. Задача конформационного поиска в таком случае сводится к нахождению наиболее «жизнеспособных» хромосом. На первом шаге генетического алгоритма случайным образом генерируется определенное количество хромосом (или, иначе говоря, конформаций). Далее, хромосомы (генотипы) проходят селекцию и ранжируются в соответствии с их «жизнеспособностью». В случае конформационного поиска это значит, что разным наборам диэдральных углов (зашифрованных в хромосомах) ставятся в соответствие конформации, в соответствии с энергиями которых производится их ранжирование. Далее вступает в силу принцип естественного отбора, который реализуется в разных вариантах. Основная сущность его состоит в том, что наиболее жизнеспособные особи (конформации с более низкой энергией) имеют больше шансов произвести потомство, что приводит к передаче потомкам и закреплению в популяции наиболее желательных признаков (значений определенных диэдральных углов) и в результате эволюции – к появлению особей с максимально возможной жизнеспособно-

стью (наиболее энергетически выгодных конформаций). При производстве потомства, кроме передачи хромосом от родителей к детям, с определенной вероятностью происходят два стохастических процесса – *мутации* и *кроссинговер*. Мутации в данном случае означают модификацию значения случайно выбранного диэдрального угла на случайную величину. Кроссинговером в биологии называется происходящий при делении клетки (точнее говоря, в мейозе) процесс обмена гомологичными участками хромосом. В данном случае это означает, что у конформаций-потомков часть диэдральных углов наследуется из одной конформации, другая – из другой. Роль мутаций и кроссинговера состоит в том, что они обогащают разнообразие популяции за счет образования новых признаков и их комбинаций. Для ускорения сходимости генетического алгоритма часто используют подходы, основанные на т.н. гибридной генетической оптимизации [157]. В этом случае особи время от времени подвергаются «*ламарковской» эволюции*[55] (в данном случае простой оптимизации геометрии до локальных минимумов энергии), и полученные таким путем геометрические характеристики молекул копируются в хромосомы и передаются потомству.

Генетический алгоритм используется в большом количестве программ. В частности, соответствующие возможности при оптимизации геометрии имеются в программах Spartan, Vega ZZ, в программах молекулярного докинга, генерации виртуальных библиотек и других. Конформационный поиск на основе генетического алгоритма реализован также в программе GAMMA [158, 159].

2.7.2.4. Методы молекулярного моделирования

Методы молекулярного моделирования объединяют 3 группы методов:

– поиск локальных минимумов на поверхности потенциальной энергии;

– генерация молекулярных траекторий (молекулярная динамика);

– стохастические методы (Монте-Карло, генетический алгоритм).

Все они основаны на вычислении энергии молекулы, как функции от пространственных координат ядер (атомов). Для расчета молекулярной энергии можно использовать как методы квантовой химии, так

[55] Ж.Б. Ламарк предполагал, что действующей силой эволюции является приспособление организмов под внешнюю среду.

и силовые поля. Наибольшее применение находят силовые поля как более быстрый метод расчета. Квантовая химия, как правило, используется только для поиска отдельных конформеров, соответствующим локальным минимумам на поверхности потенциальной энергии. В то же время квантовохимические методы иногда используются для моделирования динамики молекул. Поиск отдельных конформеров с использованием силовых полей традиционно называют *молекулярной механикой*.

Молекулярная динамика – это вычислительный метод, в котором отслеживается эволюция системы взаимодействующих атомов или частиц, происходящая в соответствии с классическими (не квантовыми) уравнениями движения (Ньютона, Лагранжа или Гамильтона). Взаимодействие между элементами системы моделируется с использованием методов силовых полей (молекулярной механики) или квантовой механики (реже). Методы молекулярной динамики и Монте-Карло не ограничены локальными минимумами и способны преодолевать энергетические барьеры между ними. В сочетании со специальными протоколами нагрева и охлаждения это может быть использовано как для перебора конформеров, так и для нахождения наиболее стабильных из них.

Метод Монте-Карло [160, 161] является одним из наиболее популярных методов стохастического поиска конформеров. На первом этапе выбирается стартовая конформация и рассчитывается ее энергия E_i. Далее эта конформация случайным образом модифицируется, например, путем изменения на случайные величины диэдральных углов некоторых связей. Для полученной конформации вычисляется энергия E_{i+1}, и она принимается за новую стартовую геометрию, если удовлетворяет *условию Метрополиса*. Условия Метрополиса можно охарактеризовать так: если $E_{i+1} \le E_i$, то конформация принимается, если же $E_{i+1} > E_i$, то конформация принимается с вероятностью $f = \exp\left(-{E_{i+1} - E_i}/{RT}\right)$, где f – фактор Больцмана. Чем больше энергия второй конформации относительно первой, тем меньше фактор Больцмана и, следовательно, тем меньше вероятность того, что вторая конформация будет принята. Таким образом, критерий Метрополиса всегда отбирает конформации с энергией ниже, чем у предыдущей, и лишь с некоторой долей вероятности (уменьшающейся с ростом энер-

гии) – конформации с более высокой энергией. Такая схема позволяет проводить отбор конформаций с низкой энергией, но при этом оставляет возможность перескока через потенциальный барьер в область другого локального минимума. Вероятность преодоления энергетического барьера, через которую возможен переход, определяется «температурой» T. Чем выше температура, тем реже отбираются устойчивые конформации, но тем меньше вероятность того, что расчет «застрянет» в одном из глубоких локальных минимумов. Метод Монте-Карло используется в пакетах MCMM [162], Flo99 (QXP) [163], модулях *obgen* и *obconformer* программы OpenBabel, реализован в конформационном модуле программы Cerius^2 компании Accelrys и многих других пакетах для молекулярного моделирования.

Алгоритм имитации отжига (англ. *simulated annealing*) – это метод решения общей задачи глобальной оптимизации (в частности, нахождения глобального минимума нелинейной функции, каковой является поверхность потенциальной энергии молекулы), основанный на имитации процесса кристаллизации вещества, происходящего при отжиге металлов. Этот алгоритм был первоначально предложен Н. Метрополисом как часть рассмотренного выше метода Монте-Карло. В последнее время его также стали использовать в сочетании с методами молекулярной динамики. В применении к задачам молекулярного моделирования этот алгоритм моделирует процесс медленного замораживания, или отжига, молекулы для того, чтобы найти ее самую стабильную конформацию. Для этого молекулу (или систему молекул), моделируемую при помощи методов Монте-Карло либо молекулярной динамики, на первом этапе нагревают до достаточно высокой температуры и дают установиться равновесию. Это позволяет молекуле преодолевать все энергетические барьеры и потенциально принимать любую конформацию. Далее система медленно охлаждается. Поскольку вероятность нахождения в области глобального минимума несколько больше, чем в области более высоко лежащих локальных минимумов, при достаточно медленном (в идеальном случае бесконечно долгом) охлаждении до температуры абсолютного нуля молекула окажется в глобальном минимуме. В реальности, однако, время охлаждения молекулы ограничено, и попадание ее в глобальный минимум не может быть гарантировано. Вследствие этого, обычно несколько раз повторяют процесс нагрева и охлаждения, в результате чего генерируется несколько наиболее выгодных конформаций молекулы. Методы

Монте-Карло и молекулярной динамики в сочетании с алгоритмом имитации отжига реализованы в достаточно большом количестве программ, в том числе в пакетах HyperChem, MOE, а также таких популярных пакетах для молекулярного моделирования, как GROMOS, NAMD и многих других.

2.7.2.5. Метод следования минимальной моде колебаний

Особый способ конформационного поиска воплощен в пакете MacroModel компании Shrödinger [164]. Его эффективный и весьма точный алгоритм поиска конформеров LMCS (или LMOD) основан на следовании минимальной моде колебаний [165]. Идея метода состоит в том, что самые низкочастотные гармонические колебания молекулы идут через самые пологие энергетические барьеры, за которыми находятся локальные минимумы, незначительно отличающиеся по энергии от исходной конформации. Следуя направлению таких колебаний, вычисляемому как собственный вектор матрицы вторых производных по координатам, отвечающий минимальному собственному значению, отыскивается ближайший локальный минимум. В нем опять рассчитывается матрица вторых производных и процедура повторяется до тех пор, пока не будут перебраны все низкоэнергетические конформеры. Энергия молекулы и ее производные вычисляются при помощи эмпирических силовых полей. Этот метод дает очень высокое качество описания биологически активных конформаций, однако требует существенных затрат времени по сравнению с другими методами [112].

2.7.2.6. Методы метрической геометрии

Методы метрической геометрии [166-168] – это математические процедуры, позволяющие генерировать декартовы координаты для набора точек исходя из матрицы расстояний между ними. Они широко распространены в молекулярном моделировании, где эффективно используются для нужд конформационного поиска. В этом случае используется обобщенная матрица расстояний, содержащая сведения о нижних и верхних пределах допустимых расстояний между каждой парой атомов. Для молекулы часть межатомных расстояний известна точно – это длины связей и расстояние между атомами, удаленными на две связи (оно может быть найдено из стандартных значений валентных углов и длин связей). В этих случаях минимальное и максимальное допустимые расстояния можно принять равными этим точно из-

вестным расстояниям. Для других, удаленных более чем на две связи пар атомов нужно указать нижний предел – обычно сумма ван-дер-ваальсовых радиусов – и верхний предел, который может быть рассчитан исходя из числа атомов в молекуле. Нижние и верхние пределы могут быть уточнены (т.е. нижние пределы подняты, а верхние – опущены), поскольку все расстояния должны удовлетворять правилу треугольника. После этого матрица расстояний инициализируется случайными числами таким образом, чтобы для каждой пары атомов это число находилось в интервале между нижним и верхним пределом. Из полученной матрицы рассчитывается метрическая матрица, содержащая в качестве элементов скалярные произведения векторов, идущих из начала координат к атомам. В том случае, если исходная матрица содержит корректный набор эвклидовых расстояний между атомами, метрическая матрица должна иметь только три ненулевых собственных значения, а соответствующие им собственные вектора – определять декартовы координаты атомов. Поскольку матрица расстояний инициализируется случайными числами, крайне маловероятно для гибких молекул, чтобы полученный таким образом набор расстояний был корректным и мог действительно соответствовать молекуле, находящейся в трехмерном физическом пространстве. Поэтому начальная конформация строится на основе собственных векторов, соответствующих трем самым большим собственным значениям. Если для этой конформации найти все межатомные расстояния, то окажется, что некоторые из них либо меньше нижнего предела, либо выше верхнего предела для данной пары атомов. Поэтому на следующем этапе производится оптимизация геометрии молекулы с использованием функции штрафа за нарушение этих пределов. В результате оптимизации будет получена конформация молекулы, удовлетворяющая всем нижним и верхним пределам.

Поскольку матрица расстояний инициализируется случайными числами, эту процедуру можно повторять множество раз, и каждый раз в общем случае будет генерироваться новая конформация, удовлетворяющая заданному набору ограничений на межатомные расстояния. Благодаря этому можно проводить сэмплинг (перебор) конформационного пространства молекул и даже межмолекулярных комплексов. Полученные конформации являются, однако, грубыми, и зачастую после дополнительной оптимизации в силовом поле молекулы попадают в один и тот же локальный минимум, однако их генерация происходит

достаточно быстро. Методы метрической геометрии дают хорошие результаты даже при наличии внутримолекулярных нековалентных связей, жестких и сшитых циклов и т.п. структурных особенностей, а также могут использоваться для систем из нескольких взаимодействующих молекул. Их недостатками являются невысокое качество геометрий систем с большими циклами и длинными цепями. Методы метрической геометрии приобрели особую популярность в конформационном поиске. Помимо применения (в некоторых случаях) в программных модулях от ChemAxon, они используются как основной метод генерации конформаций в программах DGEOM [169, 170], MOLGEO [171] (последняя используется в открытом GRID-проекте OpenMolGRID [172] для исследования и инженерии молекул), а также в сочетании с методом установки высоких искусственных потенциальных барьеров (т.н. *полинг*) [173] как один из алгоритмов конформационного поиска (т.н. *best algorithm*) в модуле ConFirm программ Cerius^2 и Catalyst [104, 150, 151].

2.7.3. Программы конвертации между представлениями

Программа, описание, сайт	**Поддерживаемые форматы и представления**	**Дополнительные опции**
OpenBabel, конвертер, http://openbabel.org (бесплатный доступ)	**Стандартные:** MOL, RXN, CML, XML, MOL2, XYZ, PDB, SMILES, SMARTS, SMIRKS, InChi, **Внешние:** очень много	2D ↔ 3D конвертация; оптимизация геометрии; генерация конформаций; расчет дескрипторов
CACTVS, графический редактор, http://www2.chemie.uni-erlangen.de/software/cactvs http://www.xemistry.com/ (бесплатный доступ)	**Стандартные:** MOL, SDF, RXN, RDF, XYZ, PDB, SMILES, SMARTS, SMIRKS, InChi, SLN **Внешние:** ChemDraw CDX, Isis Sketch SKC, ChemSketch SK2 и др.	

ACD/ChemSketch, графический редактор, http://www.acdlabs.com (коммерческий доступ, условно бесплатный (для академических целей))	**Стандартные:** MOL, SDF, RXN, SMILES, InChi, CML **Внешние:** ChemDraw CHM и CDX, Isis Sketch SKC, ChemSketch MST и RTP **Внутренний:** SK2	2D ↔ 3D конвертация; оптимизация геометрии; генерация таутомеров; предсказание logP, молярной рефракции, индекса рефракции, молярного объема, плотности; расчет формулы, состава, MW
ChemAxon (MarvinSketch, MarvinView), графический редактор, http://www.chemaxon.com/ (коммерческий доступ, условно бесплатный (для академических целей))	**Стандартные:** MOL, SDF, RXN, RDF (V2000/V3000), CML, MOL2, XYZ, PDB, SMILES, SMARTS, SMIRKS (recursive), InChi, систематические имена **Внешние:** ChemDraw CDX, Isis Sketch SKC, Gaussian CUBE, Gaussian GJF, Gamess INP **Внутренний:** MRV	2D ↔ 3D конвертация; генерация конформеров; оптимизация геометрии; молекулярная динамика; генерация таутомеров, стереоизомеров; предсказание logP, logD, pKa, распределения форм, спектров ЯМР, зарядов, поляризуемости, объема, поверхности и др.; расчет дескрипторов
Symyx Draw, графический редактор, http://accelrys.com (коммерческий доступ, условно бесплатный (для академических целей))	**Стандартные:** MOL, RXN, SMILES, InChi, IUPAC and traditional name, NEMA **Внешние:** ChemDraw CDX, Isis Sketch SKC, **Внутренний:** MOL	2D ↔ 3D конвертация (со сторонним визуализатором); генерация стереоизомеров; расчет формулы, состава, MW, биодоступности (правило 5ти), изотопомеров (масс-спектры)

OSRA: Optical Structure Recognition Application, http://cactvs.nci.nih.gov/osra/ (бесплатный доступ, с открытым кодом)	SMILES, SDF	Графические представления (GIF, JPEG, PNG, TIFF, PDF, PS) ↔ 2D конвертация
Vega ZZ, графический редактор, http://nova.colombo58.unimi.it/cms/index.php?Software_projects:VEGA_ZZ (бесплатный доступ)	**Стандартные:** MOL, MOL2, InChI, PDB, XYZ, CML, SMILES **Внешние:** Gaussian GJF, OUT, Gamess INP, Gromacs GRO и др., Tinker XYZ, Mopac ARC, DAT, HiperChem HIN и др.	2D ↔ 3D конвертация; генерация конформеров; оптимизация геометрии; молекулярная динамика (интерфейс к NAMD)
ChemOffice, графический редактор (ChemDraw), 3D редактор (Chem3D), http://www.cambridridgesoft.com (коммерческий доступ)	**Стандартные:** MOL, SDF, RXN, RDF (V2000/V3000), CML, IUPAC name **Внешние:** Isis Sketch SKC **Внутренний:** CDX	2D ↔ 3D конвертация; оптимизация геометрии; конформационный анализ (сечения ППЭ); молекулярная динамика; генерация таутомеров; предсказание logP, MP, BP, спектров ЯМР, фрагментации (масс-спектры, диссоциация), поверхности (tPSA); расчет формулы, состава, MW

2.8. ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКИХ РЕАКЦИЙ

Химическая реакция – это значительно более сложный объект для представления на компьютере, чем химическое соединение. Реакция описывается набором соединений (минимум двумя), относящимся к двум типам: реагенты и продукты. Каждая реакция происходит согласно своему механизму, описывающему процесс превращения реагентов в продукты реакции. Желательно поэтому, чтобы представление реакции как-то отражало ее механизм.

Важной задачей хемоинформатики является помощь химику в получении информации о химических реакциях, их систематизации, предсказании, обобщении и планировании синтезов. Для этого необходимо разработать методы, которые позволяют хранить, искать, сопоставлять, анализировать, сравнивать реакции, создавать модели реакционной способности, определять ограничения для реакций, анализировать цепочки, развивать методы дизайна синтеза. Оперирование химическими реакциями означает работу с их представлениями, которые могут быть созданы на разных уровнях обобщения.

Каждая реакция характеризуется набором условий ее протекания, и их изменение влияет на выход реакции, состава продуктов и даже ее направление. Следовательно, описание химической реакции должно включать как спецификацию реагентов и продуктов реакции (т.е. уравнение реакции), так и данные об условиях и реакции. Описание условий и результатов реакции может содержать информацию о температуре, давлении, растворителях, реагентах, катализаторах, выходах продуктов, регио- и стерео-селективности и других важных характеристиках реакции. Одному уравнению реакции может соответствовать множество условий.

В хемоинформатике используются различные способы описания реакций. Первый и наиболее очевидный способ представления уравнения реакции состоит в задании всех ее реагентов и продуктов, а также направления. Второй способ основан на рассмотрении реакционного центра[56] – набора атомов, меняющих свое окружение и идущих к ним связей, меняющих свой порядок, образующихся или разрывающихся в ходе химической реакции. В описание реакционного центра также включают атомы и связи, которые хотя непосредственно не задействованы в ходе реакции, но могут оказывать на него существенное влияние. Вид реакционного центра и происходящие в нем в результате реакции изменения позволяют химику отличать реакцию одного типа (например этерификации) от другого типа (например гидролиза эфира). Вследствие этого они являются ключевым элементом при кодировании и классификации химических реакций. Для корректной работы с представлениями реакций важно также указывать, как именно изменяется реакционный центр, найти точное соответствие между конкрет-

[56] Под реакционным центром иногда подразумевают отдельный атом, окружение которого меняется в результате реакции.

ными атомами и продуктов реакции. Это в ряде случаев позволяет преодолевать неоднозначности, см. пример на рис. 46.

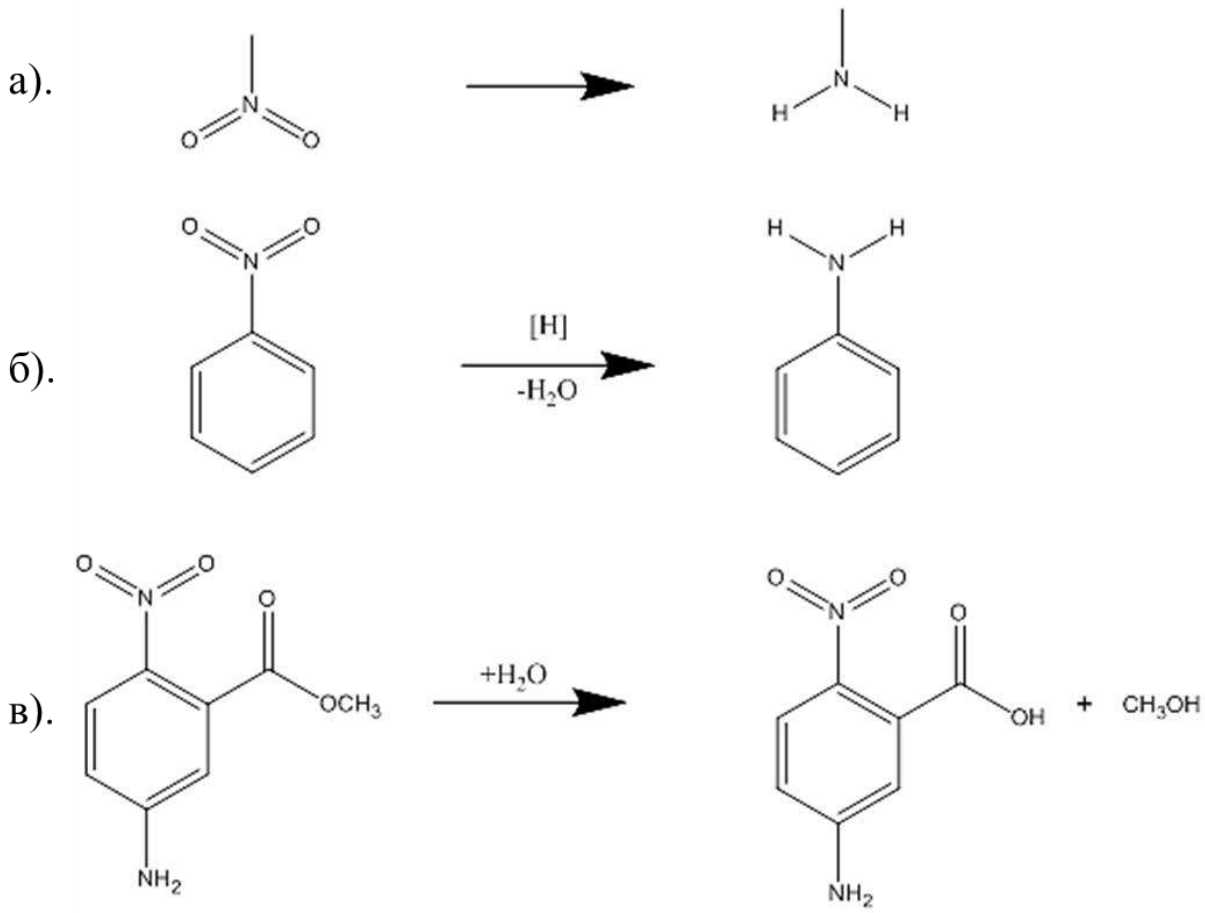


Рис. 46. Поиск с запросом (а), соответствующему восстановлению нитрогруппы, может привести к реакции (в), которая является реакцией гидролиза, а не к интересующей нас реакцией (б) восстановления

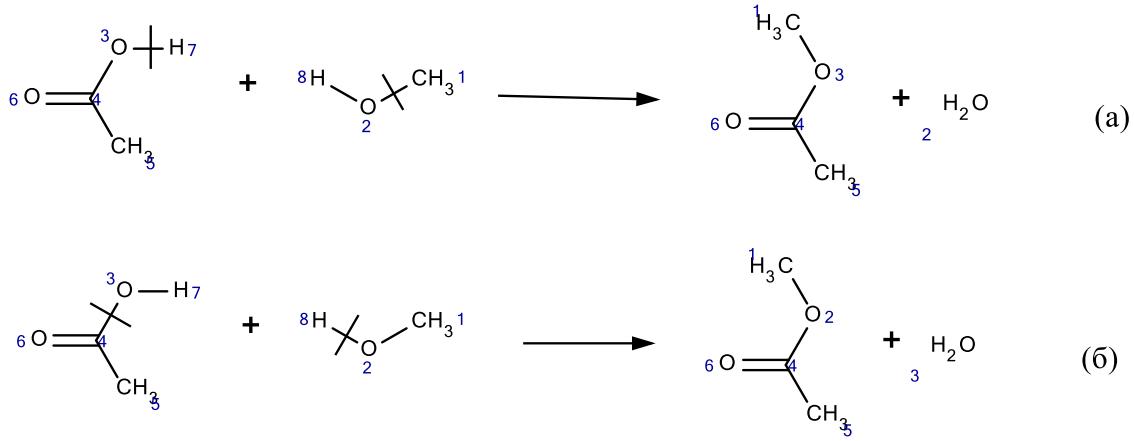


Рис. 47. Различные возможности отображения атомов в исходных веществах и продуктах

С другой стороны, точная спецификация реакционного центра может в ряде случаев представлять существенную проблему. На рис. 47 приведены две реакции, обладающие одинаковым набором реагентов и продуктов, но отличающиеся характером происходящих изменений в реакционном центре. На первый взгляд может показаться, что это одна и та же реакция. На самом же деле, реакция (а) – это присоединение метильной группы к карбоксильной, а (б) – это присоединение метоксильной группы к ацильной. Таким образом, это две разные реакции, которые поисковыми машинами будут интерпретироваться по-разному. Однако как узнать, какая из реакций действительно имеет место? В случае реакции на рис. 47 с использованием изотопных меток было показано, что правильным является уравнение (б). Правильное отображение атомов реагентов на атомы продуктов является по этой причине серьезной проблемой и в ряде случаев требует проведения специальных экспериментальных, либо теоретических исследований.

Существует несколько основных подходов для отображения атомов в реакциях. Большая часть методов использует алгоритмы отображения, основанные на поиске максимальной общей подструктуры для молекулярных [174, 175] или реберных [176] графов в реагентах и продуктах. Другой распространенный подход использует оптимизационные подходы, основанные на изоморфизме графов [177-179] или линейной целочисленной оптимизации [180, 181]. И наконец, существует алгоритм отображения, основанный на использовании генетического алгоритма [182] для оптимизации отображения атомов, базируясь на принципе «минимального химического расстояния» [183].

В качестве третьего способа представления структурных особенностей реакций можно использовать разность между определенными характеристиками продуктов и реагентов. В этом случае остается лишь то, что изменяется в результате реакции и поэтому важно для ее описания. Таким образом, по уровню обобщения можно условно разделить представления реакций на три типа:

1. Представления реакции как набора реагентов и продуктов,
2. Представления реакции как характеристик реакционного центра,
3. Представления реакции как разности продуктов и реагентов.

Иногда эти типы представлений могут пересекаться, например, представления первого типа зачастую включают элементы второго типа.

2.8.1. Представление реакции как набора реагентов и продуктов

Химики обычно представляют реакции графически при помощи уравнения: в левой его части указываются реагенты, которые отделяются друг от друга знаком сложения, далее идет стрелка, за которой следуют разделенные через знак сложения продукты реакции. Такое представление реакции можно легко закодировать для использования на компьютере путем перечисления представлений всех ее реагентов и продуктов. Этот тип представления является очень общим и может быть использован для всех видов реакций, вне зависимости от того, известен ли их механизм и приведены ли все участвующие в ней молекулы. Такое представление, однако, обладает рядом недостатков. Во-первых, поскольку реакционный центр не указан, то имеется большая неопределенность в определении даже общего типа реакции (см. рис. 46). Во-вторых, для проведения классификации реакций требуется применение сложных и неоднозначных процедур, не гарантирующих нахождение правильных решений. В-третьих, с помощью этого представления могут быть приведены неполные реакции, без указания всех реагентов и продуктов, что может сильно затруднить их анализ. Частичным решением проблемы является проведение отображения атомов реагентов на атомы продуктов, однако, для этого требуются определенные сведения о возможном механизме реакции. Для случаев, когда такой информации нет, необходимо применение одного из алгоритмов автоматического отображения атомов, которые, однако, не гарантируют правильность решения. Отображение атомов также требует, чтобы реакция была полной, с указанием всех участвующих в ней молекул, число атомов в реагентах которой должно быть равно числу атомов в ее продуктах.

Представления с указанием продуктов и реагентов с указанием отображения атомов очень широко используется в хемоинформатике. Они могут быть закодированы в линейные представления, из которых наиболее широко распространены SMILES (SMIRKS) и SLN. Форматы MDL – RXN и RDF – описывают реакции с использованием таблиц связей исходных и образующихся молекул, а также информации об отображении атомов, если она доступна. Эти форматы были описаны выше.

2.8.2. Представления реакций как характеристик реакционного центра

Данный вид представлений основан на рассмотрении строения реакционного центра и происходящих в нем (в результате реакции) изменений.

В 1963 году Г. Владуц [184] предложил проводить классификацию реакций на основании рассмотрения реакционного центра как набора атомов, при которых в ходе реакции образуются, разрушаются либо модифицируются связи. В 1986 году Г. Владуц [185] и С. Фужита [186] предложили вместо ансамбля молекулярных графов использовать один граф, метки ребер которого обозначают образование, разрушения и изменение порядков связей, для представления реакционного центра. С. Фужита расширил этот подход на представление всей реакции (а не только реакционного центра) на одном графе, названного им «*мнимым переходным состоянием*» (англ. *imaginary transition state*). Этот граф впоследствии получил название *конденсированного графа реакции* (англ. *condensed graph of reaction, CGR*). Таким образом, конденсированный граф реакции содержит *динамические* связи, которые изменяются в ходе реакции, а также *обычные* связи, остающиеся при этом неизменными (или «пар-связи» по Фужите). Динамическими связями могут быть образующиеся «*ин-связи*» и разрывающиеся «*аут-связи*». Реагенты на графе можно определить, убрав все ин-связи, а продукты – убрав все аут-связи (рис. 48). Если удалить все обычные связи и оставить только динамические связи, получится так называемый *базовый граф реакции* (англ. *basic reaction graph*).

Поскольку конденсированный граф реакции похож на молекулярный граф, для его записи можно использовать те же форматы, что и для молекул (MOL, SDF, линейный представления, фингерпринты), дополнив специальными обозначениями для динамических связей. Это дает возможность широко использовать такое представление реакций для различных целей. А. Варнек показал [187], что данный подход позволяет создать набор дескрипторов для реакций и, следовательно, проводить поиск по подобию, строить классификационные и регрессионные модели, используя подходы, разработанные для молекул. Недостатком такого представления является то, что для однозначного построения конденсированного графа реакции необходимо знание того, как отображаются атомы реагентов и в атомы продуктов.

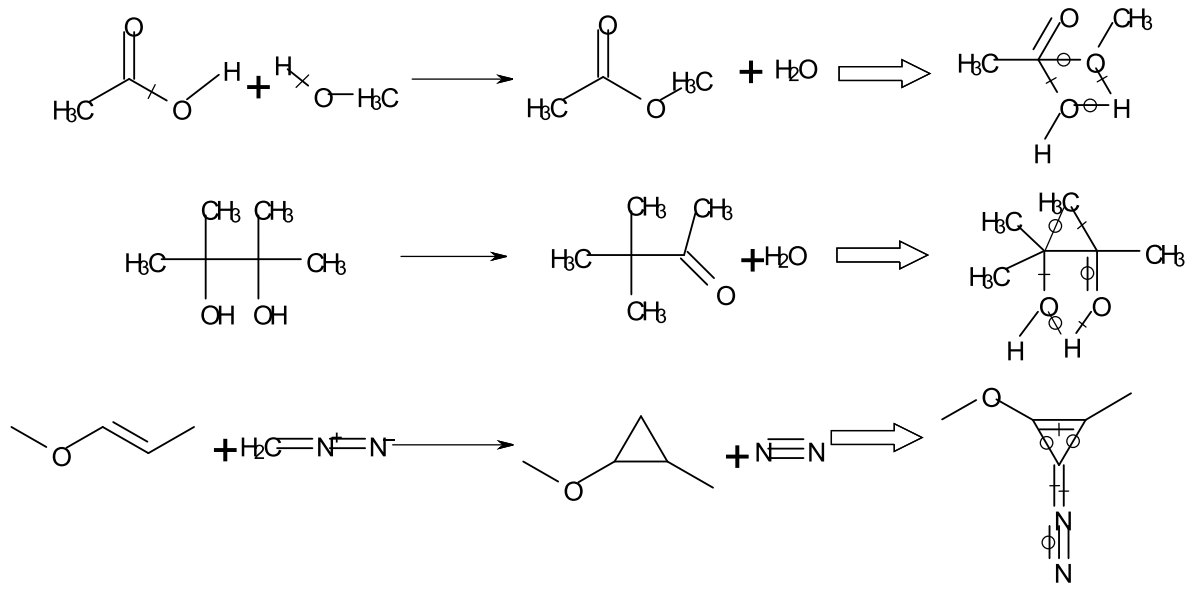


Рис. 48. Уравнения реакций и конденсированные графы реакций этерификации (сверху), пинаколиновой перегруппировки (в середине), присоединения карбена по двойной связи (снизу). Образующиеся (ин-) связи обозначены кружочком, разрывающиеся (аут-связи) зачеркнуты

Кроме вышеупомянутых, существуют и другие, менее универсальные способы представления реакций на основании знаний о строении реакционного центра и происходящих в нем изменений.

Дж. Митчел с соавторами [188] предложил представлять реакции в виде битовых строк (похожих на структурные ключи), отдельные биты которого описывают изменения, происходящие в реакционном центре. В оригинальной работе [188] предложено 58 правил, включающих рассмотрение происходящего в реакции изменения числа циклов, количества связей, вида связей (изменения порядка 1→2, 2→1 и т.д.), изменения формальных зарядов атомов, участие радикалов и т.д. Эти битовые строки могут быть использованы для тех же целей, что и молекулярные отпечатки, в частности, для поиска по сходству в базах данных, а также для кластеризации и визуализации содержащихся в них реакций. Данное представление, однако, не получило широкого распространения.

Хеш-коды реакционного центра были разработаны в компании InfoChem GmbH и использованы в алгоритме ICClassify для классификации реакций [189]. Данный подход концентрирует внимание на изменении характеристик атомов в ходе реакции. Считается, что реакционный центр образован атомами, у которых в ходе реакции изменяется число присоединенных атомов водородов, валентность, число π-

электронов, формальный заряд, либо затрагиваются идущие к ним связи. Рассмотрения реакционного центра может проходить на трех уровнях, показанных на рис. 49. На *широком* (*broad*) уровне рассматриваются только атомы реакционного центра. На *среднем* (*medium*) уровне рассматриваются как атомы реакционного центра, так и все соседние с ними тяжелые атомы (α-атомы). Наконец, на *узком* (*narrow*) уровне рассматриваются атомы реакционного центра и неводородные атомы, удаленные на 1 и 2 связи от них (α- и β-атомы), кроме sp^3-гибридного углерода. На каждом уровне, на основании характеристик атомов реакционного центра генерируется свой хеш-код (называемый ClassCode), который уникален для реакций данного типа. При рассмотрении реакций на широком уровне количество их с одинаковым хеш-кодом, как правило, больше, чем на среднем, а на среднем больше, чем на узком. Хеш-коды реакционного центра могут использоваться как для «структурной» классификации, так и для поиска сходных реакций в базах данных.

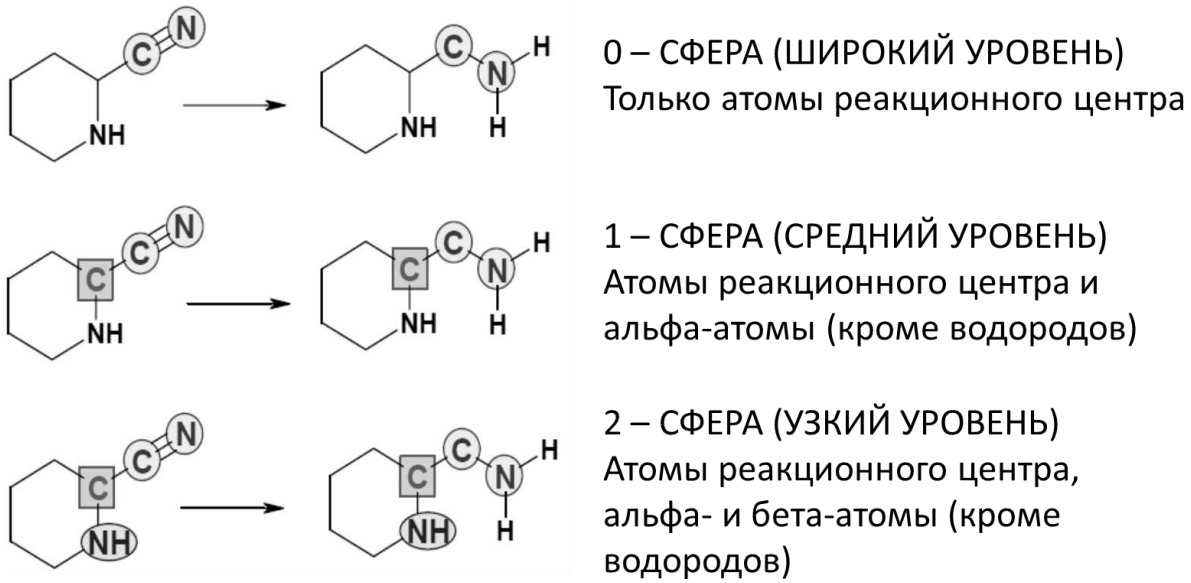


Рис. 49. Разные уровни рассмотрения реакционного центра в подходе ICClassify

Еще один подход, основанный на рассмотрении реакционного центра, был предложен в группе М. Канехисы для автоматического присвоения десятичного кода для классификации энзиматических реакций [190]. Для этого реакция разбивается на пары реагент-продукт. Далее каждый тяжелый атом в этих молекулах относят к одному из 68

типов. Сравнивая молекулярные графы для реагентов и продуктов реакции (рис. 50), находят:

– атомы, изменяющие окружение, т.е. образующие реакционный центр (R-атомы);

– атомы, которые присутствуют в продуктах реакции, но отсутствуют в реагентах (D-атомы);

– атомы, которые не меняются и не меняют свое окружение в результате превращения (M-атомы).

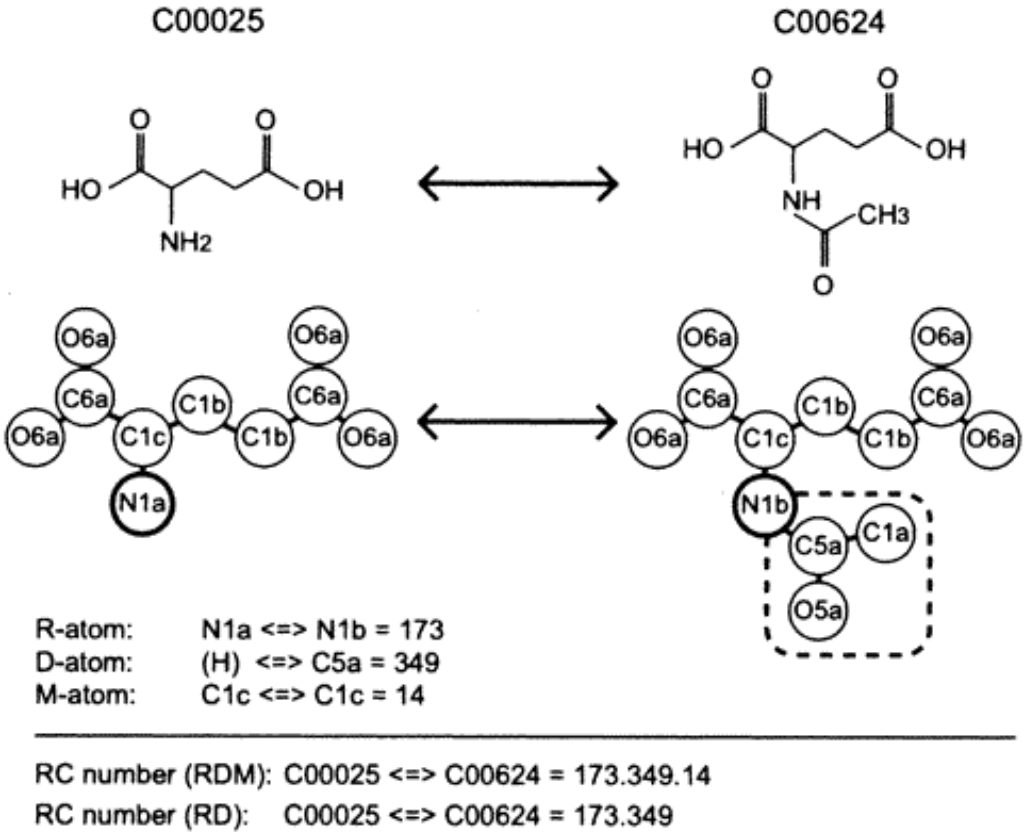


Рис. 50. Создание номеров энзиматических реакций в подходе М. Канехисы. Рисунок из работы [190] публикуется с разрешения издательства. Copyright (2004) American Chemical Society

Каждому типу изменения R-, D- и M–атомов в ходе реакции присваивается уникальный номер, который используется для создания RD- (перечислением кодов изменений R и D, как R.D) или RDM-номера реакции (перечислением кодов изменений R, D и M, как R.D.M, см. рис. 50). Этот подход был успешно использован для автоматического присвоения EC-номеров энзиматических реакций (EC – *Enzyme Commission*) [190]. Также он был применен при создании системы предсказания путей биодеградации ксенобиотиков [191].

Таким образом, рассмотрение характеристик реакционного центра может быть использовано для представления реакции практически в любом машиночитаемом виде – при помощи хэш-кода, битовой строки, таблицы связей и других.

2.8.3. Представления реакций как разности продуктов и реагентов

В середине 1970-х годов Уги и Дугунджи [192, 193] предложили описывать химические реакции с использованием так называемых R-матриц. Этот способ основан на представлении всех молекул, участвующих в реакции, в виде матрицы связей-электронов реагентов (**B**) и продуктов (**E**). Для этого требуется, чтобы количество атомов, входящих в состав реагентов, и продуктов реакции, совпадало, и каждому атому реагентов соответствовал атом в продуктах реакции, то есть проведено отображение атомов. В этом случае для описания реакции может быть использована R-матрица, определяемая как разность между **E** и **B** матрицами: $\mathbf{R} = \mathbf{E} - \mathbf{B}$. В полученной R-матрице положительные недиагональные значения будут отражать увеличение порядка связи или ее образование, отрицательные – уменьшение порядка связи либо ее разрыв, диагональные значения отражают изменение числа неподеленных электронных пар (или неспаренных электронов), рис. 51.

	1	2	3	4	5	6	7	8
1		2	1	1				
2	2				1	1		
3	1							
4	1							
5		1						
6		1						
7								1
8							1	6

B

	1	2	3	4	5	6	7	8
1		-1					1	
2	-1							1
3								
4								
5								
6								
7	1							-1
8		1					-1	

R

	1	2	3	4	5	6	7	8
1		1	1	1			1	
2	1				1	1		1
3	1							
4	1							
5		1						
6		1						
7	1							
8		1						6

E

Рис. 51. Представление Уги-Дугунджи реакции гидробромирования этилена. Для удобства нули в матрицах опущены

R-матрицы обладают следующими свойствами:

- R-матрица симметрична;

- сумма всех элементов матрицы равна 0 при сохранении числа электронов в реакции;
- сумма элементов строк или столбцов, соответствующих любому атому, также равна 0, если не меняется его формальный заряд.

Заметим, что строки или столбцы R-матрицы, соответствующие атомам, чье окружение не меняется, заполнены нулями. Удалив такие строки и столбцы, можно получить описание изменений, протекающих в реакционном центре.

Большинство элементов в R-матрице равны нулю, поэтому для экономии места ее можно модифицировать в подобие таблицы связности, перечислив только ненулевые элементы и их значения. Представление Уги-Дугунджи использовалось для множества различных целей: моделирования, классификации и предсказания новых реакций, планирования синтеза, поиска минимального расстояния между реагентом и продуктом. Недостатком данного представления также является зависимость от отображения атомов.

По-видимому, наиболее простым представлением реакций как разности продуктов и реагентов являются *разностные отпечатки реакций* (англ. *Reaction Difference Fingerprints*), предложенные компанией Daylight[57]. Для их создания используются стандартные методы формирования молекулярных отпечатков. На первом этапе создается один общий хешированный отпечаток для молекул реагентов, который отражает наличие структурных характеристик у всех реагентов в одной битовой строке, и то же самое делается для продуктов. Если структурные характеристики, использованные для получения отпечатков, реагентов, и продуктов совпадают (то есть биты одновременно равны 0 или 1 в каждом хешированном отпечатке), то в разностном реакционном отпечатке ставится 0, а если они не совпадают – то 1. Такие отпечатки могут быть использованы для оценки сходства реакций, а также их кластеризации, но не для поиска по сходству, потому что они не отражают наличие структурных фрагментов, сохраняющихся в реакции. Например, для реакции бромирования бит, соответствующий атому брома, присутствует в обоих молекулярных отпечатках, и потому будет равен нулю в разностном реакционном отпечатке. Если будет проводиться поиск реакций с участием атома брома, то данная реакция

[57] http://www.daylight.com/dayhtml/doc/theory/theory.finger.html

не выявится. Кроме того, реакционные отпечатки Daylight совпадают для прямой и обратной реакции. Достоинством данного представления является то, что для него не надо проводить отображение атомов реагентов на атомы продуктов. Главным условием его составления является стехиометричность реакции, т.е. должны быть указаны все ее реагенты и продукты.

Было также предложено несколько подходов к формированию строчного, а не битового представления реакций. Метод Риддера-Вагенера был разработан для классификации метаболических реакций [194]. С этой целью генерируются молекулярные отпечатки реагентов и продуктов на основании специального подхода, учитывающего типы атомов по правилам Sybyl. Отличием данного подхода от метода Daylight стало то, что искомые так называемые реакционные отпечатки рассчитываются как разность между молекулярными отпечатками продуктов и реагентов, и поэтому могут содержать 3 значения, равные 0, +1 или -1. В отличие от реакционных отпечатков Daylight, они не совпадают для прямой и обратной реакции.

Альтернативный подход, предложенный Ж.-Л. Фолоном с соавторами [195], представляет молекулы реагентов и продуктов в виде вектора, описывающего частоту встречаемости циркулярных подструктурных фрагментов, включающих все атомы и связи вплоть до определенного топологического расстояния от данного атома молекулы, называемых *атомными подписями* (или *сигнатурами*, англ. *signature*). Для реакции, обобщенно представляемой как $s_1S_1 + s_2S_2 + ... + s_nS_n \rightarrow p_1P_1 + p_2P_2 + ... + p_mP_m$, где s_i и p_j – стехиометрические коэффициенты перед реагентами S_i и продуктами P_j, атомная сигнатура реакции определяется как $\sigma(R) = \sum_j p_j\sigma(P_j) - \sum_i s_i\sigma(S_i)$, где σ ($P_j$) и σ ($S_i$) обозначают сигнатуры продуктов и реагентов соответственно. Это представление было использовано для автоматической классификации ферментативных реакций путем определения их EC номера. Этот вид представления не зависит от отображения атомов, однако, как и все подобные представления, не годится для подструктурного поиска.

Таким образом, разностные подходы к представлению реакций могут использоваться для разнообразных целей. Достоинствами таких представлений в большинстве случаев является независимость от отображения атомов. Тем не менее разностный характер приводит

к тому, что фрагменты, переходящие из реагентов в продукты в неизменном виде, не учитываются в представлении, что ограничивает круг решаемых с их помощью задач.

ЛИТЕРАТУРА

1. *Brown F.K.* Chemoinformatics: What is it and How does it Impact Drug Discovery / F.K. Brown // Annual Reports in Medicinal Chemistry. – 1998. – V. 33, Is. – P. 375-384.

2. G. Paris, in Extract from 218-th ACS National Meeting and Exposition New Orleans, Louisiana, August 22-26, 1999 // W.A. Warr, 1999. URL: http://www.warr.com/warrzone2000.html.

3. *Gasteiger J.* Chemoinformatics: a textbook / J. Gasteiger, T. Engel. – Weinheim: Wiley-VCH, 2003. – 649 p.

4. *Faulon J.-L.* Handbook of Chemoinformatics Algorithms / J.-L. Faulon, A. Bender. – Boca Raton: CRC Press, 2010. – 454 p.

5. *Varnek A.* Chemoinformatics as a Theoretical Chemistry Discipline / A. Varnek, I.I. Baskin // Molecular Informatics. – 2011. – V. 30, Is. 1. – P. 20-32.

6. *Vapnik V.* Statistical learning theory / V. Vapnik. – New York: Wiley, 1998. – 736 p.

7. *Lipinski C.* Navigating chemical space for biology and medicine / C. Lipinski, A. Hopkins // Nature. – 2004. – V. 432, Is. 7019. – P. 855-861.

8. *Bohacek R.S.* 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. – V. 16, Is. 1. – P. 3-50.

9. Observable universe // Wikipedia, the free encyclopedia. – URL: http://en.wikipedia.org/wiki/Observable_universe (cited 2012 December, 6).

10. *Polishchuk P.G.* 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. – V. Is. – P. 1-5.

11. *Todeschini R.* Molecular Descriptors for Chemoinformatics. Vol. I: Alphabetical listing. / R. Todeschini, V. Consonni – Weinheim: Wiley-VCH, 2009. – 967 p.

12. *Bron C.* Algorithm 457: finding all cliques of an undirected graph / C. Bron, J. Kerbosch // Commun. ACM. – 1973. – V. 16, Is. 9. – P. 575-577.

13. *Rupp M.* Graph Kernels for Molecular Similarity / M. Rupp, G. Schneider // Molecular Informatics. – 2010. – V. 29, Is. 4. – P. 266-273.

14. *Massart D.L.* Handbook of chemometrics and qualimetrics / D.L. Massart. – Amsterdam: Elsevier, 1998. – 886.

15. IUPAC Compendium of Chemical Terminology / International Union of Pure and Applied Chemistry, 2005-2012. URL: http://goldbook.iupac.org (дата обращения: 31.10.2012).

16. IUPAC Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F and H. 4th ed. / IUPAC Commission on the Nomenclature of Organic Chemistry – Oxford: Pergamon Press, 1979. – 559 p.

17. Nomenclature of Organic Compounds: Principles and Practice / Editors R.B. Fox, W.H. Powell. – Washington, DC: Oxford University Press, 2001. – 464 p.

18. *Leigh G.J.* Principles of Chemical Nomenclature: A Guide to IUPAC Recommendation / G.J. Leigh, H.A. Favre, W.V. Metanomski. – Malden: Blackwell Science, 1998. – 144 p.

19. Chemical Nomenclature and Structure Representation Division, International Union of Pure and Applied Chemistry (IUPAC), 2005-2012. URL: http://www.iupac.org/ (дата обращения: 31.10.2012).

20. *Wiswesser W.J.* A line-formula chemical notation / W.J. Wiswesser. – New York: Crowell, 1954. – 149 p.

21. *Wiswesser W.J.* How the WLN began in 1949 and how it might be in 1999 / W.J. Wiswesser // Journal of Chemical Information and Computer Sciences. – 1982. – V. 22, Is. 2. – P. 88-93.

22. *Benson F.R.* Recording and Recovering Chemical Information with Standard Tabulating Equipment / F.R. Benson // 124-th National Meeting of the American Chemical Society. – Chicago: Remington Rand Inc, 1953.

23. *Smith E.G.* Machine Searching for Chemical Structures / E.G. Smith // Science. – 1960. – V. 131, Is. 3394. – P. 142-146.

24. *Granito C.E.* Chemical Substructure Index (CSI)-A New Research Tool / C.E. Granito, M.D. Rosenberg // Journal of Chemical Documentation. – 1971. – V. 11, Is. 4. – P. 251-256.

25. *Weininger D.* SMILES. A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. / D. Weininger // Journal of Chemical Information and Computer Sciences. – 1988. – V. 28, Is. 1. – P. 31-36.

26. *Weininger D.* SMILES. 2. Algorithm for generation of unique SMILES notation / D. Weininger, A. Weininger, J.L. Weininger

// Journal of Chemical Information and Computer Sciences. – 1989. – V. 29, Is. 2. – P. 97-101.

27. *Morgan H.L.* The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service / H.L. Morgan // Journal of Chemical Documentation. – 1965. – V. 5, Is. 2. – P. 107-113.

28. *Liu X.* Computational techniques for vertex partitioning of graphs / X. Liu, K. Balasubramanian, M.E. Munk // Journal of Chemical Information and Computer Sciences. – 1990. – V. 30, Is. 3. – P. 263-269.

29. *Herndon W.C.* Canonical labeling and linear notation for chemical graphs / W.C. Herndon // Chemical Applications of Topology and Graph Theory / editor R.B. King. – Elsevier: Amsterdam, 1983. – P. 231-242.

30. *Bersohn M.* A matrix method for partitioning the atoms of a molecule into equivalence classes / M. Bersohn // Computers & Chemistry. – 1987. – V. 11, Is. 1. – P. 67-72.

31. *Uchino M.* Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. I. Vector Function for Automorphism Partitioning / M. Uchino // Journal of Chemical Information and Computer Sciences. – 1980. – V. 20, Is. 2. – P. 116-120.

32. *Hann M.* Strategic Pooling of Compounds for High-Throughput Screening / M. Hann, B. Hudson, X. Lewell // Journal of Chemical Information and Computer Sciences. – 1999. – V. 39, Is. 5. – P. 897-902.

33. *Walters W.P.* Prediction of «drug-likeness» / W.P. Walters, M.A. Murcko // Advanced Drug Delivery Reviews. – 2002. – V. 54, Is. 3. – P. 255-271.

34. *Kenny P.W.* Structure Modification in Chemical Databases / P.W. Kenny, J. Sadowski // Chemoinformatics in Drug Discovery. – Wiley-VCH Verlag GmbH & Co. KGaA, 2005. – P. 271-285.

35. *Lewell X.Q.* 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. – V. 38, Is. 3. – P. 511-522.

36. *Vieth M.* Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs / M. Vieth, M.G. Siegel, R.E. Higgs // Journal of Medicinal Chemistry. – 2003. – V. 47, Is. 1. – P. 224-232.

37. *Van Drie J.H.* ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures / J.H. Van Drie, D. Weininger, Y.C. Martin // Journal of Computer-Aided Molecular Design. – 1989. – V. 3, Is. 3. – P. 225-251.

38. *Ash S.* SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation / S. Ash, M.A. Cline, R.W. Homer // Journal of Chemical Information and Computer Sciences. – 1997. – V. 37, Is. 1. – P. 71-79.

39. *Homer R.W.* Clark. SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries / R.W. Homer, J. Swanson, R.J. Jilek, T. Hurst, R.D. Clark // Journal of Chemical Information and Modeling. – 2008. – V. 48, Is. 12. – P. 2294-2307.

40. *Трач С.С.* Комбинаторные модели и алгоритмы в химии. Лестница комбинаторных объектов и ее применение для формализации структурных задач органической химии // Принципы симметрии и системности в химии / С.С. Трач, Н.С. Зефиров. – М., 1987. – P. 54-86.

41. *Tratch S.S.* Algebraic Chirality Criteria and Their Application to Chirality Classification in Rigid Molecular Systems / S.S. Tratch, N.S. Zefirov // Journal of Chemical Information and Computer Sciences. – 1996. – V. 36, Is. 3. – P. 448-464.

42. *Granito C.E.* Computer-Generated Substructure Codes (Bit Screens) / C.E. Granito, G.T. Becker, S. Roberts // Journal of Chemical Documentation. – 1971. – V. 11, Is. 2. – P. 106-110.

43. *Durant J.L.* 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. – V. 42, Is. 6. – P. 1273-1280.

44. *Ihlenfeldt W.D.* Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility / W.D. Ihlenfeldt, Y. Takahashi, H. Abe, S. Sasaki // Journal of Chemical Information and Computer Sciences. – 1994. – V. 34, Is. 1. – P. 109-116.

45. *Tomczak J.* Data Types / J. Tomczak // Handbook of Chemoinformatics / Editor J. Gasteiger. – Weinheim: WILEY-VCH, 2003. – P. 392-409.

47. *Ihlenfeldt W.D.* Hash codes for the identification and classification of molecular structure elements / W.D. Ihlenfeldt, J. Gasteiger // Journal of Computational Chemistry. – 1994. – V. 15, Is. 8. – P. 793-813.

48. *Gasteiger J.* The WODCA System, in Software Development in Chemistry / J. Gasteiger. – Heidelberg: Springer-Verlag, 1990. – P. 57-65.

49. *Freeland R.G.* The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula / R.G. Freeland, S.A. Funk, L.J. O'Korn // Journal of Chemical Information and Computer Sciences. – 1979. – V. 19, Is. 2. – P. 94-98.

50. *Dugundji J.* An algebraic model of constitutional chemistry as a basis for chemical computer programs / J. Dugundji. – Berlin: Heidelberg: Springer. – P. 19-64.

51. *Estrada E.* Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes / E. Estrada // Journal of Chemical Information and Computer Sciences. – 1996. – V. 36, Is. 4. – P. 844 – 849.

52. *Estrada E.* Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications / E. Estrada // Journal of Chemical Information and Computer Sciences. – 1997. – V. 37, Is. 2. – P. 320-328.

53. *Estrada E.* Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 3. Molecules Containing Cycles / E. Estrada // Journal of Chemical Information and Computer Sciences. – 1998. – V. 38, Is. 1. – P. 23-27.

54. *Burden F.R.* Molecular identification number for substructure searches / F.R. Burden // Journal of Chemical Information and Computer Sciences. – 1989. – V. 29, Is. 3. – P. 225-227.

55. *Pearlman R.S.* Metric Validation and the Receptor-Relevant Subspace Concept / R.S. Pearlman, K.M. Smith // Journal of Chemical Information and Computer Sciences. – 1999. – V. 39, Is. 1. – P. 28-35.

56. *Trinajstic N.* The Laplacian matrix in chemistry / N. Trinajstic, D. Babic, S. Nikolic // Journal of Chemical Information and Computer Sciences. – 1994. – V. 34, Is. 2. – P. 368-376.

57. *Ivanciuc O.* Design of Topological Indices. Part 8. Path Matrices and Derived Molecular Graph Invariants / O. Ivanciuc,

A.T. Balaban // MATCH (Commun. Math. Chem.). – 1994. – V. 30, Is. – P. 141-152.

58. *Ivanciuc O.* Graph Theory in Chemistry / O. Ivanciuc // Handbook of Chemoinformatics. – Weinheim, 2003. – P. 103-138.

59. *Gluck D.J.* A Chemical Structure Storage and Search System Developed at Du Pont / D.J. Gluck // Journal of Chemical Documentation. – 1965. – V. 5, Is. 1. – P. 43-51.

60. *Ivanciuc O.* Canonical Numbering and Constitutional Symmetry / O. Ivanciuc // The Encyclopedia of Computational Chemistry. – Chichester: John Wiley & Sons, 1998. – P. 167-183.

61. *Berks A.H.* Markush Structure Searching in Patents/ A.H. Berks, J.M. Barnard, M.P. O'Hara // Encyclopedia of Computational Chemistry, – Chichester: John Wiley & Sons, 1998. – P. 1552-1559.

62. *Simmons E.S.* The grammar of Markush structure searching: vocabulary vs. syntax / E.S. Simmons // Journal of Chemical Information and Computer Sciences. – 1991. – V. 31, Is. 1. – P. 45-53.

63. *Barnard J.M.* A comparison of different approaches to Markush structure handling / J.M. Barnard // Journal of Chemical Information and Computer Sciences. – 1991. – V. 31, Is. 1. – P. 64-68.

64. *Пат. 1506316 А Соединенные Штаты Америки, МПК[7] С 09 В 29/366.* Pyrazolone Dye and Process of Making the Same / E.A. Markush (USA); заявитель - Pharma Chemical Corp; опубл. 26.08.1924. – 3 с.

65. *Simmons E.S.* Markush structure searching over the years / E.S. Simmons // World Patent Information. – 2003. – V. 25, Is. 3. – P. 195-202.

66. *Benichou P.* Handling Genericity in Chemical Structures Using the Markush Darc Software / P. Benichou, C. Klimczak, P. Borne // Journal of Chemical Information and Computer Sciences. – 1997. – V. 37, Is. 1. – P. 43-53.

67. *Fisanick W.* The Chemical Abstract's Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts / W. Fisanick // Journal of Chemical Information and Computer Sciences. – 1990. – V. 30, Is. 2. – P. 145-154.

68. *Ebe T.* The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 2. The MARPAT file / T. Ebe, K.A. Sanderson, P.S. Wilson // Journal of Chemical Information and Computer Sciences. – 1991. – V. 31, Is. 1. – P. 31-36.

69. *Lynch M.F.* The Sheffield Generic Structures Projecta Retrospective Review / M.F. Lynch, J.D. Holliday // Journal of Chemical Information and Computer Sciences. – 1996. – V. 36, Is. 5. – P. 930-936.

70. *Berks A.H.* Current State of Art of Markush Topological Search System / A.H. Berks // Handbook of Chemoinformatics. – Weinheim: Wiley-VCH, 2003. – P. 885-903.

71. *Engel T.* Representation of Chemical Compounds / T. Engel // Chemoinformatics. A textbook / Editors J. Gasteiger, T. Engel. – Weinheim: Wiley-VCH, 2003. – P. 15-169.

72. *Connolly M.* Analytical molecular surface calculation / M. Connolly // Journal of Applied Crystallography. – 1983. – V. 16, Is. 5. – P. 548-558.

73. *Connolly M.* Solvent-accessible surfaces of proteins and nucleic acids / M. Connolly // Science. – 1983. – V. 221, Is. 4612. – P. 709-713.

74. *Duncan B.S.* Approximation and visualization of large-scale motion of protein surfaces / B.S. Duncan, A.J. Olson // Journal of Molecular Graphics. – 1995. – V. 13, Is. 4. – P. 250-257.

75. *Lee B.* The interpretation of protein structures: Estimation of static accessibility / B. Lee, F.M. Richards // Journal of Molecular Biology. – 1971. – V. 55, Is. 3. – P. 379-IN4.

76. *Bader R.F.W.* Properties of atoms in molecules: atomic volumes / R.F.W. Bader, M.T. Carroll, J.R. Cheeseman, C. Chang // Journal of the American Chemical Society. – 1987. – V. 109, Is. 26. – P. 7968-7979.

77. Grant J.A. 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. – V. 17, Is. 14. – P. 1653-1666.

78. ROCS [Программа]. – Santa Fe, New Mexico, USA: OpenEye Scientific Software, 1997-2011.

79. *Zhokhova N.I.* Method of continuous molecular fields in the search for quantitative structure-activity relationships / N.I. Zhokhova, I.I. Baskin, D.K. Bakhronov // Doklady Chemistry. – 2009. – V. 429, Is. 1. – P. 273-276.

80. *Karpov P.V.* 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. – V. 440, Is. 2. – P. 263-265.

81. *Ritchie D.W.* Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces / D.W. Ritchie, G.J.L. Kemp // Journal of Computational Chemistry. – 1999. – V. 20. – Is. 4. – P. 383-395.

82. ParaFit [Программа]. – Bedford: CEPOS InSilico Ltd, 2009.

83. *Dalby A.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited / A. Dalby, J.G. Nourse, W.D. Hounshell // Journal of Chemical Information and Computer Sciences. – 1992. – V. 32, Is. 3. – P. 244-255.

84. *Bernstein F.C.* The protein data bank: A computer-based archival file for macromolecular structures / F.C. Bernstein, T.F. Koetzle, G.J.B. Williams // Journal of Molecular Biology. – 1977. – V. 112, Is. 3. – P. 535-542.

85. *Goff M.J.* Interpreting freezing point depression of stearic acid and methyl stearate / M.J. Goff, G.J. Suppes, M.A. Dasari // Fluid Phase Equilibria. – 2005. – V. 238, Is. 2. – P. 149-156.

86. PDB File Format Documentation, Worldwide Protein Data Bank, 2013. URL: http://www.wwpdb.org/docs.html (дата обращения: 13.10.2013).

87. *Murray-Rust P.* Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles / P. Murray-Rust, H.S. Rzepa // Journal of Chemical Information and Computer Sciences. – 1999. – V. 39, Is. 6. – P. 928-942.

88. *Murray-Rust P.* Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM / P. Murray-Rust, H.S. Rzepa // Journal of Chemical Information and Computer Sciences. – 2001. – V. 41, Is. 5. – P. 1113-1123.

89. *Gkoutos G.V.* Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust / G.V. Gkoutos, P. Murray-Rust, H.S. Rzepa, M. Wright // Journal of Chemical Information and Computer Sciences. – 2001. – V. 41, Is. 5. – P. 1124-1130.

90. *Murray-Rust P.* Chemical Markup, XML, and the World Wide Web. 4. CML Schema / P. Murray-Rust, H.S. Rzepa // Journal of Chemical Information and Computer Sciences. – 2003. – V. 43, Is. 3. – P. 757-772.

91. *Holliday G.L.* Chemical Markup, XML and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions / G.L. Holliday, P. Murray-Rust, H.S. Rzepa // Journal of Chemical Information and Modeling. – 2005. – V. 46, Is. 1. – P. 145-157.

92. *Kuhn S.* Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data / S. Kuhn, T. Helmus, R.J. Lancashire // Journal of Chemical Information and Modeling. – 2007. – V. 47, Is. 6. – P. 2015-2034.

93. IUPAC Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F and H. 4th ed. / IUPAC Commission on the Nomenclature of Organic Chemistry – Oxford: Pergamon Press, 1979. – 559 p.

94. IUPAC Provisional Recommendations. DRAFT 7 ed. – IUPAC, Chemical Nomenclature and Structure Representation Division 2004. – 1307 p. URL: http://www.iupac.org/fileadmin/user_upload/publications/recommendations/CompleteDraft.pdf (дата обращения: 11.10.2013)

95. *Wisniewski J.L.* AUTONOM: system for computer translation of structural diagrams into IUPAC-compatible names. 1. General design / J.L. Wisniewski // Journal of Chemical Information and Computer Sciences. – 1990. – V. 30, Is. 3. – P. 324-332.

96. *Goebels L.* Wisniewski. AUTONOM: system for computer translation of structural diagrams into IUPAC-compatible names. 2. Nomenclature of chains and rings / L. Goebels, A.J. Lawson, J.L. Wisniewski // Journal of Chemical Information and Computer Sciences. – 1991. – V. 31, Is. 2. – P. 216-225.

97. *Wisniewski J.L.* Autonom: Automatic Generation of IUPAC-compatiable Names from Structural Input / *J.L. Wisniewski* // Software Development in Chemistry/ Editor J. Gasteiger. – Berlin: Springer, 1990. – P. 19-29.

98. *Wisniewski J.L.* Autonom – A Chemist's Dream: System for (Micro) Computer Generation of IUPAC-Compatiable Names from Structural Input / J.L. Wisniewski // Chemical Structures / Editor W.A. Warr. – Berlin: Springer, 1993. – P. 55-63.

99. *Wisniewski J.L.* Autonom: A Computer Program for Generation of IUPA Systematic Nomenclature Directly from the Graphys Structure Input / J.L. Wisniewski // Recent Advances in Chemical Information / Editor H. Collier. – Cambridge: Royal Society of Chemistry, 1993. – P. 77-87.

100. ChemAxon Kft., Záhony u. 7, Building HX, 1031 Budapest, Hungary. 2012; URL: http://www.chemaxon.com.

101. CambridgeSoft, 100 CambridgePark Drive, Cambridge, MA 02140 USA. 2012; URL: http://www.cambridgesoft.com.

102. Advanced Chemistry Development, Inc., 8 King Street East, Suite 107, Toronto, Ontario M5C 1B5, Canada 2012; URL: http://www.acdlabs.com.

103. ChemInnovation Software, Inc., 7966 Arjons Dr., Suite A, San Diego, CA 92126, USA 2012; URL: http://cheminnovation.com.

104. Accelrys, Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA. 2012; URL: http://accelrys.com.

105. Bio-Rad Laboratories, 1000 Alfred Nobel Drive, Hercules, CA 94547, USA. 2012; URL: http://www.bio-rad.com.

106. *Taylor R.* Life-science applications of the Cambridge Structural Database / R. Taylor // Acta Crystallographica Section D. – 2002. – V. 58, Is. 6 Part 1. – P. 879-888.

107. *Allen F.H.* Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry / F.H. Allen, W.D.S. Motherwell // Acta Crystallographica Section B. – 2002. – V. 58, Is. 3 Part 1. – P. 407-422.

108. *Allen F.H.* The development of versions 3 and 4 of the Cambridge Structural Database System / F.H. Allen, J.E. Davies, J.J. Galloy // Journal of Chemical Information and Computer Sciences. – 1991. – V. 31, Is. 2. – P. 187-204.

109. *Allen F.* The Cambridge Structural Database: a quarter of a million crystal structures and rising / F. Allen // Acta Crystallographica Section B. – 2002. – V. 58, Is. 3 Part 1. – P. 380-388.

110. *Berman H.M.* The Protein Data Bank / H.M. Berman, J. Westbrook, Z. Feng // Nucleic Acids Research. – 2000. – V. 28, Is. 1. – P. 235-242.

111. *Sadowski J.* Representation of 3D Structures / J. Sadowski // Handbook of Chemoinformatics: From Data to Knowledge / Editor J. Gasteiger. – Weinheim: WILEY-VCH, 2003. – P. 231-261.

112. *Schwab C.F.* Conformation Analysis and Searching / C.F. Schwab // Handbook of Chemoinformatics: From Data to Knowledge / Editor J. Gasteiger. – Weinheim: WILEY-VCH, 2003. – P. 262-301.

113. *Leach A.R.* The application of Artificial Intelligence to the conformational analysis of strained molecules / A.R. Leach, K. Prout,

D.P. Dolata // Journal of Computational Chemistry. – 1990. – V. 11, Is. 6. – P. 680-693.

114. *Leach A.R.* Automated conformational analysis: Algorithms for the efficient construction of low-energy conformations / A.R. Leach, K. Prout, D.P. Dolata // Journal of Computer-Aided Molecular Design. – 1990. – V. 4, Is. 3. – P. 271-282.

115. *Leach A.R.* An investigation into the construction of molecular models by the template joining method / A.R. Leach, K. Prout, D.P. Dolata // Journal of Computer-Aided Molecular Design. – 1988. – V. 2, Is. 2. – P. 107-123.

116. *Dolata D.P.* WIZARD: AI in conformational analysis / D.P. Dolata, A.R. Leach, K. Prout // Journal of Computer-Aided Molecular Design. – 1987. – V. 1, Is. 1. – P. 73-85.

117. *Dolata D.P.* WIZARD: applications of expert system techniques to conformational analysis. 1. The basic algorithms exemplified on simple hydrocarbons / D.P. Dolata, R.E. Carter // Journal of Chemical Information and Computer Sciences. – 1987. – V. 27, Is. 1. – P. 36-47.

118. *Leach A.R.* Automated conformational analysis: Directed conformational search using the A* algorithm / A.R. Leach, K. Prout // Journal of Computational Chemistry. – 1990. – V. 11, Is. 10. – P. 1193-1205.

119. *Leach A.R.* A combined model-building and distance-geometry approach to automated conformational analysis and search / A.R. Leach, A.S. Smellie // Journal of Chemical Information and Computer Sciences. – 1992. – V. 32, Is. 4. – P. 379-385.

120. *Leach A.R.* An Algorithm To Directly Identify a Molecule's «Most Different» Conformations / A.R. Leach // Journal of Chemical Information and Computer Sciences. – 1994. – V. 34, Is. 3. – P. 661-670.

121. *Pearlman R.S.* 3D Molecular Structures: Generation and Use in 3D Searching, in 3D QSAR / *R.S. Pearlman* // Drug Design: Theory, Methods and Applications / editor H. Kubinyi. – Escom: Leiden, 1993. – P. 41-79.

122. *Pearlman R.S.* Rapid Generation of High Quality Approximate 3D Molecular Structures / R.S. Pearlman // Chem. Des. Auto. News. – 1987. – V. 2, Is. – P. 1-7.

123. *Sadowski J.* From atoms and bonds to three-dimensional atomic coordinates: automatic model builders / J. Sadowski, J. Gasteiger // Chemical Reviews. – 1993. – V. 93, Is. 7. – P. 2567-2581.

124. *Sadowski J.* Three-dimensional Structure Generation: Automation / J. Sadowski // Encyclopedia of Computational Chemistry. – Chichester: John Wiley & Sons, 1998. – P. 2976-2988.

125. *Gasteiger J.* Automatic generation of 3D-atomic coordinates for organic molecules / J. Gasteiger, C. Rudolph, J. Sadowski // Tetrahedron Computer Methodology. – 1990. – V. 3, Is. 6, Part C. – P. 537-547.

126. *Sadowski J.* The generation of 3D models of host-guest complexes / J. Sadowski, C. Rudolph, J. Gasteiger // Analytica Chimica Acta. – 1992. – V. 265, Is. 2. – P. 233-241.

127. *Dale J.* Exploratory Calculations of Medium and Large Rings. Part 1. Conformational Minima of Cycloalkanes / J. Dale // Acta Chem. Scand. – 1973. – V. 27, Is. – P. 1115-1129.

128. *Klebe G.* A fast and efficient method to generate biologically relevant conformations / G. Klebe, T. Mietzner // Journal of Computer-Aided Molecular Design. – 1994. – V. 8, Is. 5. – P. 583-606.

129. *Klebe G.* Methodological developments and strategies for a fast flexible superposition of drug-size molecules / G. Klebe, T. Mietzner, F. Weber // Journal of Computer-Aided Molecular Design. – 1999. – V. 13, Is. 1. – P. 35-49.

130. *Schwab C.H.* Konformative Flexibilitaet von Liganden im Wirkstoffdesign: PhD Thesis / C.H. Schwab. – Erlangen, 2001. – 264 p.

131. OMEGA. Conformer Ensembles Containing Bioactive Conformations [Программа] – Santa Fe, New Mexico, USA: OpenEye Scientific Software, 1997-2012.

132. Bostrom J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations / J. Bostrom, J.R. Greenwood, J. Gottfries // Journal of Molecular Graphics and Modelling. – 2003. – V. 21, Is. 5. – P. 449-462.

133. *Perola E.* Conformational Analysis of Drug-Like Molecules Bound to Proteins. An Extensive Study of Ligand Reorganization upon Binding / E. Perola, P.S. Charifson // Journal of Medicinal Chemistry. – 2004. – V. 47, Is. 10. – P. 2499-2510.

134. *Mayo S.L.* Goddard. DREIDING: a generic force field for molecular simulations / S.L. Mayo, B.D. Olafson, W.A. Goddard // The Journal of Physical Chemistry. – 1990. – V. 94, Is. 26. – P. 8897-8909.

135. *Halgren T.A.* Merck molecular force field. IV. conformational energies and geometries for MMFF94 / T.A. Halgren, R.B. Na-

chbar // Journal of Computational Chemistry. – 1996. – V. 17, Is. 5-6. – P. 587-615.

136. *Halgren T.A.* Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules / T.A. Halgren // Journal of Computational Chemistry. – 1996. – V. 17, Is. 5-6. – P. 616-641.

137. *Halgren T.A.* Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94 / T.A. Halgren // Journal of Computational Chemistry. – 1996. – V. 17, Is. 5-6. – P. 553-586.

138. *Halgren T.A.* Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 / T.A. Halgren // Journal of Computational Chemistry. – 1996. – V. 17, Is. 5-6. – P. 490-519.

139. *Halgren T.A.* Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions / T.A. Halgren // Journal of Computational Chemistry. – 1996. – V. 17, Is. 5-6. – P. 520-552.

140. *Wipke W.T.* Analogy and Intelligence in Molecular Model Building / W.T. Wipke // Artificial Intelligence Applications in Chemistry / ed. T. Pierce and B. Hohne. – Washington: American Chemical Society, 1986. – P. 136-146.

141. *Wipke W.T.* AIMB: Analogy and intelligence in model building. System description and performance characteristics / W.T. Wipke, M.A. Hahn // Tetrahedron Computer Methodology. – 1988. – V. 1, Is. 2. – P. 141-167.

142. *Wipke W.T.* Chemical Structures / W.A. Warr. – Berlin: Springer, 1988. – P. 267-268.

143. *Hahn M.A.* Chemical Structures / M.A. Hahn. – Berlin: Springer, 1988. – P. 269-278.

144. *Davies K.* Experiences building and searching the Chapman & Hall Dictionary of Drugs/ K. Davies, R. Upton // Three-dimensional chemical structure handling / Editors Y.C. Martin, P. Willett. – Amsterdam: Elsevier, 1990. – P. 665-671.

145. *Lipton M.* The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space / M. Lipton, W.C. Still // Journal of Computational Chemistry. – 1988. – V. 9, Is. 4. – P. 343-355.

146. *Goto H.* Corner flapping: a simple and fast algorithm for exhaustive generation of ring conformations / H. Goto, E. Osawa // Journal of the American Chemical Society. – 1989. – V. 111, Is. 24. – P. 8950-8951.

147. *Smellie A.* 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. – V. 35, Is. 2. – P. 285-294.

148. *Smellie A.* Analysis of Conformational Coverage. 2. Applications of Conformational Models / A. Smellie, S.D. Kahn, S.L. Teig // Journal of Chemical Information and Computer Sciences. – 1995. – V. 35, Is. 2. – P. 295-304.

149. P.W. Sprague. Automated chemical hypothesis generation and database searching with Catalyst / P.W. Sprague // Perspect. Drug Discov. Des. – 1995. – V. 3, Is. – P. 1-20.

150. Catalyst [Программа], доступна в составе продуктов компании Accelrys. San Diego: Accelrys, 2012.

151. Cerius 2 [Программа], доступна в составе продуктов компании Accelrys. San Diego: Accelrys, 2012.

152. *Goldberg D.E.* Genetic Algorithms in Search, Optimization, and Machine Learning. 1989, New York: Addison-Wesley Professional. 432.

153. *Dorigo M.* Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique / M. Dorigo, M. Birattari, T. Stützle // IEEE Computational Intelligence Magazine. – 2006. – V. 1, Is. 4. – P. 28-39.

154. *Shmygelska A.* An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem / A. Shmygelska, H. Hoos // BMC Bioinformatics. – 2005. – V. 6, Is. 1. – P. 30.

155. *Korb O.* PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design / O. Korb, T. Stützle, T.E. Exner //Ant Colony Optimization and Swarm Intelligence, 5th International Workshop, ANTS 2006 / Editors M. Dorigo, L.M. Gambardella, M. Birattari, A. Martinoli, R. Poli, T. Stützle. – Heidelberg: Springer-Verlag, 2006. – P. 247-258.

156. *Bonabeau E.* Swarm Intelligence: From Natural to Artificial Systems. / E. Bonabeau, G. Theraulaz, M. Dorig. – Oxford: Oxford University Press, 1999. – 320 p.

157. *Nair N.* Genetic Algorithms in Conformational Analysis / N. Nair, J.M. Goodman // Journal of Chemical Information and Computer Sciences. – 1998. – V. 38, Is. 2. – P. 317-320.

158. *Handschuh S.* Superposition of Three-Dimensional Chemical Structures Allowing for Conformational Flexibility by a Hybrid Method / S. Handschuh, M. Wagener, J. Gasteiger // Journal of Chemical Information and Computer Sciences. – 1998. – V. 38, Is. 2. – P. 220-232.

159. *Handschuh S.* The Search for the Spatial and Electronic Requirements of a Drug / S. Handschuh, J. Gasteiger // Journal of Molecular Modeling. – 2000. – V. 6, Is. 2. – P. 358-378.

160. *Metropolis N.* Equation of State Calculations by Fast Computing Machines / N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller // The Journal of Chemical Physics. – 1953. – V. 21, Is. 6. – P. 1087-1092.

161. *Metropolis N.* The Monte Carlo Method / N. Metropolis, S. Ulam // Journal of the American Statistical Association. – 1949. – V. 44, Is. 247. – P. 335-341.

162. *Chang G.* An internal-coordinate Monte Carlo method for searching conformational space / G. Chang, W.C. Guida, W.C. Still // Journal of the American Chemical Society. – 1989. – V. 111, Is. 12. – P. 4379-4386.

163. *McMartin C.* QXP: Powerful, rapid computer algorithms for structure-based drug design / C. McMartin, R.S. Bohacek // Journal of Computer-Aided Molecular Design. – 1997. – V. 11, Is. 4. – P. 333-344.

164. Schrödinger, Inc., 120 West 45th Street, Tower 45, New York, NY 10036-4041, USA. 2012; URL: http://www.schrodinger.com.

165. *Kolossvary I.*Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides / I. Kolossvary, W.C. Guida // Journal of the American Chemical Society. – 1996. – V. 118, Is. 21. – P. 5011-5019.

166. *Crippen G.M.* Distance Geometry and Molecular Conformations/ G.M. Crippen, T.F. Havel // Chemometrics Research Studies Series 15 / Editor D. Bawden. – Taunton: Research Studies Press, 1988. – 541 p.

167. *Crippen G.M.* Global energy minimization by rotational energy embedding / G.M. Crippen, T.F. Havel // Journal of Chemical Information and Computer Sciences. – 1990. – V. 30, Is. 3. – P. 222-227.

168. *Wenger J.C.* Deriving three-dimensional representations of molecular structure from connection tables augmented with configuration designations using distance geometry / J.C. Wenger, D.H. Smith // Journal of Chemical Information and Computer Sciences. – 1982. – V. 22, Is. 1. – P. 29-34.

169. DGEOM 95: Distance Geometry [Программа], QCPE Program No. 590/ J.M. Blaney, G.M. Crippen, A. Dearing, J.S. Dixon. – Columbus, Ohio, USA: Computational Chemistry List, Ltd., 1995.

170. *Blaney J.M.* Distance Geometry in Molecular Modelling / J.M. Blaney, J.S. Dixon // Reviews in Computational Chemistry / Editors K.B. Lipkowitz, D.B. Boyd. – New York: Wiley-VCH, 1994. – P. 299-335.

171. *Gordeeva E.V.* Rapid conversion of molecular graphs to three-dimensional representation using the MOLGEO program / E.V. Gordeeva, A.R. Katritzky, V.V. Shcherbukhin // Journal of Chemical Information and Computer Sciences. – 1993. – V. 33, Is. 1. – P. 102-111.

172. *Sild S.* Open Computing Grid for Molecular Science and Engineering / S. Sild, U. Maran, A. Lomaka, M. Karelson // Journal of Chemical Information and Modeling. – 2006. – V. 46, Is. 3. – P. 953-959.

173. *Smellie A.* Poling: Promoting conformational variation / A. Smellie, S.L. Teig, P. Towbin // Journal of Computational Chemistry. – 1995. – V. 16, Is. 2. – P. 171-187.

174. *Funatsu K.* Automatic recognition of reaction site in organic chemical reactions / K. Funatsu, T. Endo, N. Kotera // Tetrahedron Computer Methodology. – 1988. – V. 1, Is. 1. – P. 53-69.

175. *McGregor J.J.* Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions / J.J. McGregor, P. Willett // Journal of Chemical Information and Computer Sciences. – 1981. – V. 21, Is. 3. – P. 137-140.

176. *Korner R.* Automatic Determination of Reaction Mappings and Reaction Center Information. 1. The Imaginary Transition State Energy Approach / R. Korner, J. Apostolakis // Journal of Chemical Information and Modeling. – 2008. – V. 48, Is. 6. – P. 1181-1189.

177. *Akutsu T.* Efficient Extraction of Mapping Rules of Atoms from Enzymatic Reaction Data / T. Akutsu // Journal of Computational Biology. – 2004. – V. 11, Is. 2-3. – P. 449-462.

178. *Crabtree J.D.* Automated reaction mapping / J.D. Crabtree, D.P. Mehta // J. Exp. Algorithmics. – 2009. – V. 13, Is. – P. 1.15-1.29.

179. *Heinonen M.* Computing atom mappings for biochemical reactions without subgraph isomorphism / M. Heinonen, S. Lappalainen, T. Mielikainen // J Comput Biol. – 2011. – V. 18, Is. 1. – P. 43-58.

180. *First E.L.* Stereochemically Consistent Reaction Mapping and Identification of Multiple Reaction Mechanisms through Integer Linear Optimization / E.L. First, C.E. Gounaris, C.A. Floudas // Journal of Chemical Information and Modeling. – 2012. – V. 52, Is. 1. – P. 84-92.

181. *Latendresse M.* Accurate Atom-Mapping Computation for Biochemical Reactions / M. Latendresse, J.P. Malerich, M. Travers // Journal of Chemical Information and Modeling. – 2012. – V. Is. –

182. *Fontain E.* The problem of atom-to-atom mapping. An application of genetic algorithms / E. Fontain // Analytica Chimica Acta. – 1992. – V. 265, Is. 2. – P. 227-232.

183. *Jochum C.* The Principle of Minimum Chemical Distance (PMCD) / C. Jochum, J. Gasteiger, I. Ugi // Angewandte Chemie International Edition in English. – 1980. – V. 19, Is. 7. – P. 495-505.

184. *Vladutz G.E.* Concerning one system of classification and codification of organic reactions / G.E. Vladutz // Information Storage and Retrieval. – 1963. – V. 1, Is. 2-3. – P. 117-146.

185. *G. Vladutz,* Do we still need a classification of reactions? / G.E. Vladutz // Modern Approaches to Chemical Reaction Searching / Editor P. Willett. – London: Gower, 1986. – P. 202-220.

186. *Fujita S.* Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts / S. Fujita // Journal of Chemical Information and Computer Sciences. – 1986. – V. 26, Is. 4. – P. 205-212.

187. *Varnek A.* Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures / A. Varnek, D. Fourches, F. Hoonakker // Journal of Computer-Aided Molecular Design. – 2005. – V. 19, Is. 9. – P. 693-703.

188. *O'Boyle N.M.* Using Reaction Mechanism to Measure Enzyme Similarity / N.M. O'Boyle, G.L. Holliday, D.E. Almonacid, J.B.O. Mitchell // Journal of Molecular Biology. – 2007. – V. 368, Is. 5. – P. 1484-1499.

189. ICClassify. The InfoChem Reaction Classification Program [Программа] – Gröbenzell, Germany: InfoChem GmbH, 2009.

190. *Kotera M.* Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions / M. Kotera, Y. Okuno, M. Hattori // Journal of the American Chemical Society. – 2004. – V. 126, Is. 50. – P. 16487-16498.

191. *Oh M.* Systematic Analysis of Enzyme-Catalyzed Reaction Patterns and Prediction of Microbial Biodegradation Pathways / M. Oh, T. Yamada, M. Hattori // Journal of Chemical Information and Modeling. – 2007. – V. 47, Is. 4. – P. 1702-1712.

192. *Dugundji J.* An algebraic model of constitutional chemistry as a basis for chemical computer programs / J. Dugundji, I. Ugi // Computers in Chemistry. – Heidelberg: Springer, 1973. – P. 19-64.

193. *Ugi I.* Chemical similarity, chemical distance, and computer-assisted formalized reasoning by analogy/ I. Ugi, M. Wochner, E. Fontain, J. Bauer, B. Gruber, R. Karl // Concepts and Applications of Chemical Similarity / Editors M.A. Johnson, G.M. Maggiora. – New York: Wiley, 1990. – P. 239-288.

194. *Ridder L.* SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites / L. Ridder, M. Wagener // ChemMedChem. – 2008. – V. 3, Is. 5. – P. 821-832.

195. *Faulon J.-L.* Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor / J.-L. Faulon, M. Misra, S. Martin // Bioinformatics. – 2008. – V. 24, Is. 2. – P. 225-233.

СОДЕРЖАНИЕ


МОЛЕКУЛА
ПРЕДСТАВЛЕНИЕ
ХИМИЧЕСКИЙ
АТОМ
СТРУКТУРА
ДАННЫЕ
МОЛЕКУЛЯРНЫЙ
СОЕДИНЕНИЕ
РЕАКЦИЯ
СВЯЗЬ
СТРОКА
ЧИСЛО
ПРОГРАММА
ПОВЕРХНОСТЬ
КОНФОРМАЦИЯ
ХЕМОИНФОРМАТИКА
ГРАФ
ИНФОРМАЦИЯ
МЕТОД
ФРАГМЕНТ
ИСПОЛЬЗОВАНИЕ
МАТРИЦА
ОБЪЕКТ
СВОЙСТВО
SMILES
ПОИСК
НАЗВАНИЕ
ОПИСАНИЕ
ЗНАЧЕНИЕ
АЛГОРИТМ
СЛУЧАЙ
ТАБЛИЦА
ОБРАЗ
СИСТЕМА
БИТ