РОССИЙСКАЯ АКАДЕМИЯ НАУК

Институт Органической Химии имени Н.Д.Зелинского

На правах рукописи

Кумсков Михаил Иванович

МЕТОДОЛОГИЯ ПРОГНОЗИРОВАНИЯ СВОЙСТВ ХИМИЧЕСКИХ СОЕДИНЕНИЙ И ЕЕ ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

05.13.17. - теоретические основы информатики

Диссертация на соискание ученой степени
доктора физико-математических наук

Москва- 1997

Содержание

0. Введение.....5

1. Глава 1. Методы и программы установления зависимостей структура-свойство химических веществ.....13

1.1 Квантовохимические методы. 14

1.2 Анализа пространственных соответствий в системе “рецептор-лиганд”.....19

1.3 Построение QSAR-моделей - количественных корреляций "структура - биологическая активность".....24

1.3.1 Параметрическое описание молекул.....28

1.3.2 QSAR-анализ трехмерных молекул 29

1.3.3 Трехмерные дескрипторы молекул.....32

1.3.4 Подструктурные молекулярные дескрипторы. 35

1.4 Построение количественных корреляций "структура- физико-химическое свойство".....43

1.4.1 Аддитивные подходы оценки физико-химических свойств веществ.....43

1.4.2 Теоретико-графовые индексы молекулярных структур.....46

1.4.3 Использование мер подобия молекул. 53

1.5 Выводы 54

2. Глава 2. Формирование структурных дескрипторов молекулярных графов.....56

2.1 Задача построения количественных корреляций "структура-свойство" как задача распознавания образов. 58

2.2 Структурные дескрипторы и структурные спектры.....66

2.2.1 Формирование матрицы "молекула-признак" (QSAR-матрицы). 69

2.2.2 Маркирование атомов.....70

2.2.3 Свойства структурных спектров маркированных графов 79

2.3 Несвязные фрагменты и их представление.....81

2.3.1 Примитивы описания молекулярных графов 81
2.3.2 Кодирование несвязных фрагментов..... 82
2.4 Алгоритмы анализа структур молекулярных графов 85
2.4.1 “Волновой” алгоритм и матрица расстояний графа..... 85
2.4.2 Связность графа..... 88
2.4.3 Алгоритмы работы с кольцевыми системами М-графа. 90
2.5 Выводы. 98

3. Глава 3. Применение эволюционных алгоритмов для решения QSAR-задач 99
3.1 Метод Группового Учета Аргументов (МГУА) 99
3.1.1 Общая схема МГУА 100
3.1.2 Модифицированная схема построения семейств QSAR-уравнений 103
3.2 МГУА-поиск сложных фрагментов, адекватных свойству..... 107
3.2.1 Обобщенные фрагменты М-графов. 107
3.2.2 Составные фрагменты М-графов. 109
3.2.3 Относительные фрагменты М-графа. 111
3.3 Генерация молекулярных графов с использованием эволюционных алгоритмов 112
3.3.1 Переборное решение обратной QSAR-задачи..... 112
3.3.2 Схема эволюционной генерации молекулярных графов..... 113
3.3.3 Алгоритм эволюционной генерации..... 114
3.4 Выбор метрик М-графов с помощью генетического алгоритма.... 119
3.4.1 Схема отбора метрики 120
3.4.2 Генетический алгоритм 120
3.5 Выводы 124

4. Глава 4. Программная система для конструирования QSAR-моделей..... 125
4.1 Основные типы данных и программные модули QSAR-системы. 125
4.2 Объектно-ориентированый подход к представлению структур данных в QSAR-системе. 129
4.2.1 Компьютерное представление М-графа. 129

4.2.2 Компьютерное представление представление обучающей БД. 135
4.2.3 Компьютерное представление МГУА-данных. 137
4.3 Программная система BIBIGON. 138
4.3.1 Базовые маркеры атомов 139
4.3.2 Работа пользователя с системой 140
4.4 Программная система ChemAdd - развитие системы BIBIGON. 145
4.4.1 Редактор обучающей выборки химических структур. 145
4.4.2 Массовая генерация и селекции графов как метод решения обратной QSAR-задачи. 147
4.4.3 Проблема отказа от прогноза свойства сгенерированной структуры. 149
4.5 Организация базы знаний, надстраиваемой над структурными БД. 151
4.6 Выводы 154

5. Глава 5. Прогнозирование физико-химических свойств веществ. 155
5.1 Свойства алканов. 156
5.2 Энтальпия образования веществ разных химических классов. 167
5.3 Поляризуемость молекул. 169
5.4 Температура кипения для фуранов/тетрафуранов и тиофенов 171
5.5 Липофильность замещенных бензолов. 173
5.6 Чувствительность к удару нитросоединений. 175
5.7 Температура плавления нитросоединений. 177
5.8 Хроматографическая подвижность антрациклиновых антибиотиков ряда даунорубицина. 178
5.9 Анализ большой СБД по липофильности. 180
5.10 Анализ большой СБД по ^{13}C спектрам 182
5.11 Выводы 186

6. Глава 6. Прогнозирование биологической активности. 187
6.1 Использование трехмерного представления молекул 187
6.1.1 Антибактериальная активность триазинов. 187

6.1.2 Анти-ВИЧ активность..... 191
6.2 Использование поверхности электростатического потенциала для формирования структурных дескрипторов..... 193
6.2.1 Общая схема формирования дескрипторов, описывающих особенности пространственной формы молекул 196
6.2.2 Мускусный запах..... 203
6.2.3 Психотропная активность. 206
6.3 Выводы 209

7. Заключение..... 210

8. Литература..... 218

0. Введение

Задача нахождения зависимостей (количественных корреляций) “структура-свойство” химических соединений привлекает в последнее время повышенное внимание в связи с развитием вычислительной техники, дешевых накопителей большого объема и соответственно быстрым ростом структурных химических баз данных (СБД). Широкое распространение компьютерных коммуникаций в сети INTERNET привело к оснащению рабочих мест химиков-исследователей мощными персональными ЭВМ, которые имеют оперативный доступ к фактографическим структурным базам данных (СБД), содержащим сведения о свойствах и структурную химическую информацию Возникает естественное желание использовать уже имеющиеся данные “структура молекулы - проявляемые свойства” для *проведения обобщений* с тем, чтобы иметь возможность целенаправленно проводить внеэкспериментальный поиск новых соединений, обладающих заданным набором свойств. Такой поиск новых веществ является актуальным во многих прикладных химических исследованиях, включая ограническую химию, агрохимию, медицинскую химию и фармакологию.

Разработка методов поиска количественных корреляций "структура-свойство" (QSAR) химических соединений (QSAR - "Quatitative Structure-Activity Relationship") выделилась в отдельное научное направление. QSAR-модели используются для предварительной внеэкспериментальной оценки свойств новых соединений. QSAR-моделирование позволяет проводить формализацию и обобщение дорогостоящих экспериментальных данных, накапливаемых в компьютерных СБД, получать ответы на вопросы о том, в каком направлении следует планировать синтез новых соединений. В настоящее время существует ряд отечественных и зарубежных QSAR систем. Как правило, они используют заранее предопределенный и фиксированный набор молекулярных дескрипторов (или дескрипторных центров), на основании которых строятся сложные (составные) признаки и проводится поиск QSAR зависимостей. Вычисляемые признаки ориентированы разработчиками системы на конкретный вид свойства. Ключевым аспектом проведения QSAR-моделирования является выбор описания молекул в виде

вектора признаков. Адекватный выбор признаков позволяет содержательно проводить оценку найденных QSAR-моделей.

В работе развивается принципиально новый подход к QSAR моделированию. Признаковые пространства, в которых проводится описание молекул СБД, формируются автоматически под управлением эксперта-химика с последовательным усложнением детализации описания и уровня представления молекул. Тип признакового пространства определяется уровнем сложности представления структуры в ЭВМ. Структура молекулы может быть представлена на нескольких уровнях детализации: на топологическом, на двумерно-топологическом (с указанием планарной проекции атомов), в трехмерном виде в минимальной энергетической конформации, в трехмерном виде с дополнительным расчетом пространственных электростатических потенциалов. Заранее не известно, на каком уровне следует проводить описание молекул (находящихся в СБД только в виде таблиц связи атомов) для QSAR-анализа конкретного свойства. Выбор уровня представления молекул и адаптация признаков для конкретного свойства проводится экспертом в процессе построения и селекции QSAR моделей. Таким образом, увеличение предсказательной силы QSAR зависимостей проводится не за счет использования специальных классификационных моделей (нечеткие множества, нейронные сети, иерархические классификаторы) или усложнения вида функциональной QSAR зависимости, а на основе постепенной детализации описаний молекул, которые последовательно представляются на различных уровнях.

Известно, что топологические и информационные индексы молекулярных графов могут быть представлены в виде функций от их структурных дескрипторов. Они являются универсальными и легко интерпре тируемыми признаками химических соединений. Это дало возможность широко использовать их в аддитивных схемах расчета. Построение аддитивных схем расчета физико-химических свойств веществ является нетривиальной проблемой. Поэтому весьма актуальна автоматизация решения этой задачи и исследование границ применения QSAR-моделей, построенных на структурных дескрипторах различных типов.

Работа выполнена в соответствии с планом научных работ Института Органической Химии РАН по теме **43**: "Разработка алгебраических и теоретико-

графовых моделей для решения структурных задач органической химии" (номер государственной регистрации **01880006030**), а также в рамках следующих проектов Российского фонда фундаментальных исследований (РФФИ):

93-012-1045: "Унифицированные математические модели и программно-инструментальные системы для прогнозирования новых органических соединений с заданными свойствами";

94-01-00041: "Инструментальная система формирования баз знаний о зависимостях "структура-свойство" органических соединений на основе символьного представления фрагментов молекулярных графов";

96-01-01598: "Распознавание пространственных форм молекул биологически активных соединений с целью компьютерного предсказания свойств новых веществ".

97-07-90307: “Селекция метрик для поиска подобных молекул в структурных фактографических БД с использованием знаний “структура-свойство”.

Целью работы является разработка унифицированной методологии для решения QSAR-задач прогнозирования свойств химических соединений на основе поиска адекватного представления молекул в виде структурных спектров и использования эволюционного программирования, а также построение на этой основе семейств прогностичных QSAR-моделей для ряда физико-химических и биологически свойств.

В работе были поставлены следующие задачи:

1.Провести анализ существующих программных QSAR систем с точки зрения используемых методов описания молекул и методов поиска QSAR-зависимостей.

2.Обосновать использование структурных спектров в качестве единообразного способа описания молекул для решения QSAR-задач.

3.Провести формализацию основных этапов компьютерного QSAR-моделирования и определить функциональную организацию программной системы. Разработать структуры данных для представления в ЭВМ ее основных объектов.

4.Разработать алгоритмы и программы анализа структур молекулярных графов и автоматического формирования структурных спектров молекул органических соединений различных классов.

5.Разработать эволюционные алгоритмы и программы для поиска структурных спектров, адекватных для исследуемого свойства, и для построения семейств QSAR-уравнений.

6.В рамках разработанной QSAR-системы провести вычислительные эксперименты по поиску зависимостей “структура-свойство” для различных химических классов веществ и различных свойств, сравнить полученные результаты с известными литературными данными.

Научная новизна работы состоит в следующем.

1.Впервые предложена и практически исследована общая методика построения QSAR-зависимостей на основе поиска адекватных признаковых пространств молекул, автоматически порождаемых под управлением эксперта. Основу методики составляют:

- Концепция многоуровневого представления молекулы как структурного объекта;
- Экспертная классификация локальных свойств атомов и определение “особых точек” (или базовых фрагментов) - примитивов описания молекул;
- Автоматическое индуктивное порождение структурных спектров заданной сложности на основе выбранных примитивов молекул;
- Использование эволюционных алгоритмов - метода группового учета аргументов (МГУА), - для выбора типа структурного спектра, адекватно характеризующего свойство молекул, и для построения семейств QSAR зависимостей, образующих модель.

2.Впервые разработана и программно реализована методика QSAR-оценки свойств органических соединений на основе формирования базы знаний "структура-свойство", запросом к которой является молекулярный граф.

3.Создан объектно-ориентированный программный инструментарий для быстрого конструирования программных QSAR-систем для ПЭВМ типа IBM

PC. На его основе разработана система BIBIGON и ее расширенная версия - система ChemAdd. Системы позволяют проводить на ЭВМ полный цикл работ по QSAR-моделированию.

4.Выполнено систематическое исследование прогностической устойчивости QSAR-моделей, посторенных на различных типах структурных спектров для ряда физико-химических и биологических свойств. Найденые QSAR-модели не уступают, а в ряде случаев и превосходят по точности известные QSAR зависимости.

Практическая ценность результатов. QSAR-системы BIBIGON и ChemAdd позволяют автоматизировать построение и селекцию аддитивных схем расчета многих физико-химических свойств органических соединений.

QSAR-системы BIBIGON и ChemAdd используются в научных институтах РАН и РАМН для проведения QSARмоделирования и поиска новых соединений с заданным профилем свойств.

Разработанный программный инструментарий может быть использован для быстрого создания прототипов программ анализа молекулярных графов, вычисления специализированных маркеров атомов, автоматического формирования новых индуктивных структурных спектров, для построения и тестирования соответствующих QSAR-зависимостей.

Найденные прогностически устойчивые QSAR-модели, хранящиеся в базе знаний системы, могут использоваться для проведения внеэкспериментального скрининга новых органических соединений и предсказания их свойств.

Апробация работы. Результаты работы докладывались:

- на 8-ой Всесоюзной конференции "Использование ЭВМ в спектроскопии молекул и химических исследованиях" (*Новосибирск*, **1989**);
- на Всесоюзной научной конференции "Оценка фармакологической активности химических соединений. Принципы и подходы" (*Купавна*,**1989**);
- на Межвузовской конференции "Молекулярные графы в химических исследованиях" (*Калинин*, **1990**);

•на Всесоюзной школе-семинаре "Статистический и дискретный анализ данных и экспертные оценки" (*Одесса*, **1991**);

•на 1-ой Всесоюзной конференции по теоретической органической химии ВАТОХ (*Волгоград*, **1991**);

•на 9-ой Всесоюзной конференции "Химическая информатика" (Ч*ерноголовка*, **1992**);

•на 1-ой Международной конференции "Химия, технология и применение фторсодержащих соединений в промышленности" (*Санкт-Петербург*, **1994**);

•на 2-ом Российском конгрессе "Человек и лекарство" (*Москва*, **1995**);

•на 21-ом Международном семинаре по пиротехнике (*Москва*, **1995**);

•на 2-ой Всероссийской конференции "Распознавание образов и обработка изображений РОАИ-2-95" (*Ульяновск*, **1995**);

•на Координационном совещании секции "Информатика" Научного совета по комплексной проблеме "Кибернетика" РАН (*Ульяновск*, **1995**);

•на 7-ой Всероссийской конференции "Математические методы распознавания образов" (*Пущино*, **1995**);

•на 5-ом Международном симпозиуме по молекулярным аспектам химиотерапии (*Гданьск*, **1995**);

•на 9-ом Европейском симпозиуме по органической химии (*Варшава*, **1995**); на Первом съезде фармакологов Украины (*Полтава*, **1995**);

•на 4-ом Открытом германо-российском семинаре "Распознавание образов и понимание изображений" (*Валдай*, **1996**);

•на Первой международной конференции по применению эволюционных вычислений и их приложениям EvCA'96 (*Москва*, **1996**);

•на научных семинарах академика Н.С.Зефирова (*Москва*, **1988-1997**);

•на научных семинарах по теории графов профессора А.А.Зыкова (*Одесса*, **1990,1991,1993**);

•на научных семинарах профессора Ю.Г.Папулова на кафедре физической химии Тверского госуниверситета (*Тверь*, **1993-1995**);

•на научно-исследовательском семинаре академика Н.С.Бахвалова на кафедре вычислительной математики механико-математического факультета МГУ (*Москва*, **1994,1997**);

•на научно-исследовательском семинаре профессора М.Р.Шура-Бура "Автоматизация программирования" на факультете Вычислительной математики и кибернетики МГУ (*Москва,* **1994,1995, 1997**);

•на научно-исследовательском семинаре академика Ю.И.Журавлева в Вычислительном центре РАН (*Москва,***1994**).

Работа состоит из введения, шести глав, заключения, списка литературы и двух приложений.

В *первой главе* дается обзор современных подходов к установлению взаимосвязи структура-активность с использованием вычислительной техники. Рассмотрена функциональная организация и собенности нескольких программных QSAR-систем, а также общие принципы организации программной системы, предназначенной для проведения полного цикла работ по прогнозированию свойств химических соединений и поиску новых соединений с заданными свойствами. Проведена классификация существующих способов описания молекулярных химических графов (МХ-графов) на основе топологических и теоретико-информационных индексов, а так же с использованием подструктур молекулярного графа. Рассмотрена общая схема QSAR-адаптации эволюционного метода отбора набора адекватных дескрипторов - метода группового учета аргументов (МГУА).

Вторая и третья главы являются центральными главами работы. Во второй главе сформулирована общая постановка задачи установления количественных корреляций структура-свойство (QSAR) как специальный случай задачи распознавания образов. Рассмотрен метод решение задачи выбора набора дескрипторов для описания молекулярных графов обучающей выборки, наиболее адекватного исследуемому свойству. Рассматрены алгоритмы анализа МХ-графов.

В *третьей главе* описано применение эволюционных алгоритмов для решения различных этапов QSAR-задач. Рассматривается решение “обратной QSAR-задачи”, т.е. задачи поиска химических веществ, обладающих заданными свойствами, на основе массовой генерации и селекции MX-графов.

В *четвертой главе* рассматриваются принципы построения QSAR-систем, структуры данных и программные объекты, реализующие описанные выше алгоритмы и методики, а также две программные QSAR-системы BIBIGON и ChemAdd, реализованные в рамках созданных библиотек объектов.

В *пятой главе* приведены практические результаты, полученные при использовании созданных программных систем для прогнозирования физико-химических свойств веществи иллюстрирующие эффективность методики автоматического выбора адекватного описания структур для QSAR-задач.

В *шестой главе* разработанная методика успешно применена для прогнозирования таких биологических свойств как антикокцидная активность, анти-ВИЧ активность, психотропная активность и амбровый запах.

Заключение содержит основные выводы работы. В *приложении 1* приведен графический вид молекулярных структур для ряда описанных в 5-ой и 6-ой главах QSAR-моделей. В *приложении 2* подробно описан пользовательский интерфейс QSAR-системы BIBIGON.

Основные результаты диссертации опубликованы в работах, список которых приведен в конце заключения.

1. Глава 1. Методы и программы установления зависимостей структура-свойство химических веществ.

Одной из научных проблем, решение которой имеет важное значение для современных технологий, является построение моделей, связывающих строение и свойства веществ. В настоящее время разработка методов поиска количественных корреляций "структура-свойство" (QSAR - "Quantitative Structure-Activity Relationship") выделилась в отдельное научное направление. Решение QSAR-задач позволяет без синтеза молекул некоторого химического ряда прогнозировать их свойства по заданным характеристикам строения. QSAR-модели используются для предварительной внеэкспериментальной оценки свойств новых соединений. QSAR-моделирование позволяет проводить обобщение дорогостоящих экспериментальных данных, накапливаемых в компьютерных структурных базах данных (СБД), и получать ответы на вопросы о том, в каком направлении следует планировать синтез новых соединений.

Как правило, при построении QSAR-моделей используется заранее предопределенный набор признаков (дескрипторов), описывающий структурные особенности молекул. Признаки, выбираемые химиком-экспертом, ориентированы на анализ конкретного свойства в заданном химическом ряду соединений. Выбор описания молекул в виде вектора признаков является ключевым моментом проведения QSAR-моделирования, поскольку этот выбор сводит задачу построения QSAR уравнения *к классичесой задаче распознавания образов со стандартной информацией* [1,2,3].

QSAR задача имеют все специфические черты задачи распознавания образов [4]. Для этих задач трудно строить формальные теории и применять классические математические методы. Уровень формализации предметной области - химии, - таков, что он не может составить основу для синтеза модели, отвечающей классическим математическим канонам и допускающей изучение аналитическими или численными методами. Математическая модель может быть построена для оценки ряда важных свойств молекул - на этом подходе основаны *квантово-химические модели,* обсуждаемые ниже. Однако, синтез и изучение таких моделей связано с существенными

вычислительными затратами, которые в ряде случает существенно превышают выигрыш, приносимый решением задачи.

Современные подходы к установлению взаимосвязи структура-свойство/активность можно условно отнести к одной из трех категорий:

1.Подходы, основанные на квантовохимических расчетах, на решении и оптимизации систем волновых уравнений.

2.Подходы, основанные на использовании трехмерного моделирования исследуемых соединений для анализа пространственных соответствий в системе “рецептор-лиганд”.

3.Классический QSAR подход (построение количественных корреляций структура-свойство - ККСС) - поиск таких различий в структуре соединений обучающей СБД, которые влияют на изменение значений исследуемого свойства/активности молекул .

Первые два подхода основаны на построении математических моделей анализируемого явления и не относятся к QSAR задачам. Ниже они кратко охарактеризованы для сохранения общности описания известных методов оценки свойств веществ

1.1 Квантовохимические методы.

Квантовохимические методы обычно используют для решения задач расчета теоретических параметров молекул, таких как потенциалы ионизации, индексы реакционной способности, теплоты образования, поверхности потенциальной энергии и переходных состояний [5, 6], и определение устойчивых конформаций молекул [7].

Квантовая механика и статистическая физика позволяют дать исчерпывающее объяснение любым экспериментальным данным о таком важном свойстве как реакционная способность органического соединения, а также предсказать возможные направления реакций [8]. Однако для этого необходимо располагать очень мощными ЭВМ и достаточно точными вычислительными методами. За последние годы в этих областях был достигнут существенный прогресс [9].

В основе современной квантовой химии лежит уравнение Шрёдингера для стационарных состояний [10]. Электронное строение молекулы как системы электронов в кулоновском поле ядер, пространственное расположение которых известно, можно определить из ее многоэлектронной волновой функции Ψ_k, зависящей от координат всех электронов. Волновые функции Ψ_k и энергии E_k стационарных электронных состояний системы определяются как решения независящего от времени уравнения Шредингера:

$$H \Psi_k = E_k \Psi_k \qquad (1.1)$$

где H - полный электронный гамильтониан системы: $H = T_e + V_{ee} + V_{en}$. Оператор T_e отвечает кинетической энергии электронов, оператор V_{ee} описывает кулоновское взаимодействие между электронами, а оператор V_{en} - между электронами и ядрами. Волновая функция Ψ_k должна быть однозначной, непрерывной, обращаться в нуль на бесконечности, иметь интегрируемый квадрат модуля и удовлетворять принципу Паули. Физический смысл волновой функции Ψ_k состоит в том, что ее квадрат описывает распределение электронной плотности в молекуле, когда молекула находится в состоянии с номером **k**. Индекс **k** нумерует собственные функции оператора Гамильтона в порядке возрастания энергии системы. Основное состояние молекулярной системы отвечает минимальному значению полной электронной энергии.

Уравнения Шредингера обычно решают в предположении, что ядерную и электронную волновые функции можно разделить и решать уравнения для движения ядер и электронов раздельно. Многоэлектронную волновую функцию записывают в виде детерминанта, построенного из одноэлектронных волновых функций φ_μ, называемых спин-орбиталями. Волновые функции φ_μ являются произведением функции ψ_μ, зависящей только от пространственного положения электрона, и спиновой функции, принимающей одно из двух (1/2, -1/2) значений. Функция ψ_μ называется орбиталью. Для атомов это будет атомная орбиталь (АО), а для молекул - молекулярная орбиталь (МО). Для большинства органических молекул в основном состоянии все спин-орби-

тали разбиваются на пары с одинаковой пространственной частью и противоположными спинами электронов, т.е. на каждой МО располагается два электрона. явный вид орбитали находится из уравнения Хартри-Фока:

$$F\psi_\mu = e_\mu \psi_\mu \quad \mu = 1,2,... \quad (1.2)$$

Оператор Хартри-Фока ***F*** включает оператор кинетической энергии одного электрона, притяжение электрона к ядрам и члены, описывающие суммарное действие на данный электрон кулоновского потенциала всех остальных электронов системы и обменное взаимодействие. Уравнения Хартри-Фока (1.2) в численной форме были решены для всех основных атомов и ионов и в результате был определен вид атомных орбиталей и их энергии. Расчет МО многоатомных молекул непосредственно из уравнения (1.2) является весьма трудоемкой задачей по сравнению с расчетом АО атомов. В связи с этим для изучения электронного строения многоатомных молекулярных систем МО разлагают по базису АО:

$$\psi_\mu = \sum_{i=1,n} (a_{\mu i} \chi_i) \qquad \mu = 1,2, ..., n; \qquad (1.3)$$

где χ_i - атомные орбитали, вид которых известен, ***n*** - число включенных в базис АО, $a_{\mu i}$ - коэффициенты, определяющие вид АО χ_i в МО ψ_μ. В формуле (1.3) МО аппроксимируются линейными комбинациями АО (ЛКАО). Существуют и другие методы расчета МО, подробное описание теории МО можно найти в целом ряде монографий [10, 11]. Однако теория строения органических соединений развивается почти исключительно в рамках метода ЛКАО, который позволяет связать характеристики электронного строения и реакционную способность соединений с параметрами электронных спектров атомов и межатомных взаимодействий. С физической точки зрения представление МО в виде ЛКАО означает предположение, что в окрестности каждого атома, где влияние потенциала более далеких атомов ослабевает, одноэлектронная волновая функция должна походить на атомную.

В зависимости от подходов к решению уравнения (1.3) расчетные квантовохимические методы делятся на *неэмпирические (ab initio)* и *полуэмпирические*. В первых проводится строгое решение уравнения (1.3) с учетом всех электронов молекулы и расчетом всех интегралов межмолекулярного взаимодействия. С ростом размеров структуры сложность расчетов

возрастает пропорционально *четвертой степени* числа базисных АО, что затрудняет ab initio расчет больших молекул. Важным моментом неэмпирических расчетом является выбор базисных АО, по которым разлагается МО, т.к. точность расчета физических свойств молекул зависит от полноты базиса. Так, для расчета геометрии, теплот образования, дипольных моментов, электростатических потенциалов и орбитальных энергий органических молекул, состоящих из небольшого числа атомов (около 10), обычно достаточно использовать минимальный базис. Однако, для расчета энергий изомеризации, возбужденных состояний молекул и заряженных ионов требуются расширенные базисные наборы с включением высоколежащих атомных состояний. Неэмпирические методы являются *приближенными*, прежде всего из-за неполноты использованного базиса. Большинство неэмпирических расчетов проводят в базисах небольшого и среднего размеров. Это может вносить существенную ошибку в результаты расчета, так как такие базисные наборы не могут передать некоторые особенности распределения электронной плотности в молекулах. В результате все без исключения параметры молекул вычисляются с некоторой ошибкой.

Общая особенность всех *полуэмпирических методов* состоит в упрощенной записи выражения для матричных элементов оператора Хартри-Фока за счет пренебрежения многоцентровыми интегралами межэлектронного взаимодействия. Чтобы компенсировать ошибку, вносимую в запись матричных элементов, вводят эмпирические параметры - характеристики атомов и межатомных взаимодействий, которые определяют на основе экспериментальных данных для нескольких соединений, выбираемых в качестве стандарта. Благодаря быстрому развитию квантовой химии были разработаны достаточно эффективные полуэмпирические методы, которые можно использовать для изучения реакционной способности больших молекул, представляющих интерес не только для органической химии, но и для биохимии[12,13,14]. Так, возможно установить, какие факторы определяют направление и относительный выход продуктов реакции, а также получить недоступную для эксперимента информацию о геометрии и электронной структуре переходных состояний.

На основе анализа приближений, сделанных при разработке того или иного квантовохимического метода, как правило, *не удается установить*

область его применения и очертить круг задач, которые можно решить с его помощью. К сожалению, многие квантовохимические методы, которые лучше обоснованы с теоретической точки зрения, на практике дают плохие результаты и поэтому не применяются, а более грубые модели с удачно подобранными параметрами широко используются [8]. Это связано с тем, что в любом квантовохимическом методе сделано достаточно много различных приближений. В некоторых методах ошибки, к которым приводят эти приближения, частично компенсируют друг друга и в результате получается хорошее согласие с экспериментом. Сказать заранее, будет или не будет иметь место такая компенсация, нельзя, поэтому выяснить область применения и охарактеризовать точность каждого конкретного метода можно лишь на основе численного эксперимента и систематизации опубликованного расчетного материала [9]. Приближенный характер квантовохимических расчетов приводит к тому, что для вычисления различных параметров молекул приходится пользоваться разными методами, и правильный *выбор их является важной частью работы* [15]. Лишь при рациональном и обоснованном выборе метода квантовохимические расчеты позволяют получить интересную информацию о реакционной способности органических соединений, дать объяснение экспериментально наблюдаемым закономерностям и даже сделать некоторые предсказания. При этом следует помнить, что получить *надежный* численный результат для каждой конкретной реакции квантовая химия в настоящее время не позволяет [8].

С помощью квантовохимических расчетов можно получить много интересной информации о строении, реакционной способности, физико-химических свойствах и биологической активности органических соединений [16]. Методы квантовой химии позволяют расчитывать параметры любых молекул, интермедиатов и даже переходных состояний. Поэтому в настоящее время их широко применяют в химических исследованиях [17]. Ежегодно публикуется более тысячи работ в этой области. С помощью квантовохимических методов расчитываются параметры, характеризующие электронную структуру молекулы. Исходя из этих параметров, можно определить важные свойства молекулы: способность молекулы отдавать или принимать электрон, энергию образования комплекса с другой молекулой, термодинамические параметры, реакционную способность и т.п. Одно из

приложений квантовохимических методов состоит в расчете возможных устойчивых конформаций. Устойчивые конформации (т.е. отвечающие наименьшей энергии) находятся путем расчета зависимости энергии молекул от вращения связей при фиксированных значениях длин и углов связей. Методами квантовой механики было исследовано множество биологически активных соединений: антималярийные агенты, анестизирующие средства, транквилизаторы, анальгетики, гербициды, пестициды и другие [14,16,18].

1.2 Анализа пространственных соответствий в системе "рецептор-лиганд".

Подходы, связанные с геометрическим моделированием взаимодействия биологически активных молекул ("*лиганда*") с белковой макромолекулой ("*рецептором*") можно разделить на две группы в зависимости от того, известна или нет пространственная структура активного центра (binding site) рецептора [19]. Образно говоря, эти подходы активно используют метафору "ключ-замок", где ключом является лиганд - биологически активная молекула. Замок - активный центр рецептора, - представляет собой "карман" или полость, содержащую "активные области" - функциональные группы со специфическими локальными физико-химическими свойствами. Для того, чтобы молекула могла быть "ключом", во-первых, ее размеры не должны превышать отверстия "кармана", а во-вторых, пространственное расположение функциональных групп "ключа" должно комплементарно соответствовать "активным областям замка" - лиганд должен быть аффинным рецептору (affinity). На современном этапе развитие метафоры "ключ-замок" привело к формулировке концепции *молекулярного распознавания* [20] и интенсивно разрабатывающемуся на практике понятию молекулярного подобия [256].

Методы, использующие известную структуру рецептора. Существует значительное число программных пакетов для конструирования лекарств на этой основе [21, 22, 23]. В работе [24] дан обзор различных методов, которые были успешно применены для конструирования молекул при известном активном центре рецептора. Основное их назначение состоит в предсказании аффиности - комплементарного сходства активного центра рецептора и конструируемого программой лиганда с целью его последующего

приоритетного химического синтеза. Была предложена штрафная функция общего вида, позволяющая оценить аффиность по известной трехмерной структуре рецептора и лиганда. В настоящее время большинство методов для оценки аффинности основываются на вычислении энергии связывания, которые включают вычисление водородных связей, оценку энергии вращения вокруг связей и учет гидрофобных взаимодействий [21]. Энергия связывания ΔG вычисляется как сумма по всем атомам **a** лиганда, попавшим в карман **r** рецептора:

$$\Delta G = \sum_{r=1}^{n_s} \sum_{a \in r} \sum_{i=1}^{n_p} (p_{a,i} \cdot \varepsilon_{r,i}) + CE$$

где n_s - число активных центров (карманов) рецептора; **a** - атом лиганда, попавший в карман **r**; n_p - число физико-химических параметров, описывающих атом **a**; $p_{a,i}$ - ***i***-ое значение физико-химического параметра атома **a**; $\varepsilon_{r,i}$ - подбираемый коэффициент, характеризующий ***i***-ый параметр в кармане **r**; CE - параметр, характеризующий компоненту конформационной энергии.

Для конструирования (генерации) новых структур лигандов (при известной пространственной модели активного центра рецептора) используют два различных подхода. В *первом подходе* используется библиотека заранее заданных молекулярных фрагментов. Из нее выбираются фрагменты, которые состыковываются (dock) с активными участками кармана рецептора как *изолированные элементы*, и затем соединяются дополнительными атомами друг с другом так, чтобы сформировать молекулярную структуру. Эта стратегия генерации используется, например, в программе GROW [25] и в программе LUDI [21]. Во *втором подходе* проводится "выращивание" (генерация) структуры лиганда внутри рецепторного кармана путем последовательного присоединения при каждой итерации дополнительного атома или фрагмента. В программе GROUPBUILD [24] лиганд "выращивается" путем добавления фрагмента из библиотеки фрагментов. На каждом шаге генерации оцениваются все возможные добавления библиотечных фрагментов путем расчета энергии методом молекулярной механики. Другая программа LEGEND [26] использует атомы для сборки лиганда так, чтобы не

нарушались стерические (Вандерваальсовы) ограничения при заполнении рецепторного кармана [27]. Затем вызывается дополнительная программа LORE для отбора структур с малой энергией (оцениваемой методом молекулярной механики) и правильным расположением водородных связей с функциональными группами рецептора. Этот же метод поатомного "выращивания" используется в программе GrowMol [22].

Методы, восстанавливающие 3D-модель рецептора. Если активный центр рецептора неизвестен, то возможно построение его модели на основе знания пространственных конформаций биологически активных молекул, содержащихся в обучающей БД. Основная трудность при решении этой задачи состоит в том, что с одной стороны, гибкие молекулы имеют несколько устойчивых конформаций, а с другой стороны, при связывании с рецептором энергия структуры лиганда, как правило, не соответствует глобальному энергетическому минимуму [28, 29]. Разработано ряд подходов, позволяющих успешно решить эту проблему [30, 31], которая получила название построение *количественной модели активного центра (quantitative binding site model)* [32]. Особенность пространственного строения активного центра часто представляют в виде *модели фармакофора* как пространственное расположение нескольких функциональных групп лиганда (обладающих заданными локальными физико-химическими свойствами), которые существенны для распознавания и активизации исследуемого рецептора [33,34]. Начало развития компьютерной методологии поиска моделей фармакофоров в рамках концепции "ключ-замок" относится еще к началу 80-х годов [35]. Программы, реализующие ряд подходов [31,36], включены в мощную систему молекулярного моделирования SYBYL американской фирмы Tripos [37], с ее помощью получены модели фармакофоров для различных типов биологической активности [38, 39].

В 1988 году Doweyko предложил метод HASL (Hypotetical Active Site Lattice) для построения модели фармакофора в виде ортогональной сетки, каждой точке которой присваивается усредненное значение исследуемой активности [40]. Активность новой структуры оценивается на основе сопоставления HASL сетки рецептора и сетки, наложенной на исследуемую молекулу. При этом используется итеративная процедура для оптимизации пространственного расположения сеток. Эта оригинальная и идейно очень

простая модель была успешно применена для исследования ряда биологических активностей [41, 42].

Один из перспективных и мощных методов построения модели рецептора реализован в алгоритме Compass [32, 43, 44]. Его основные отличия от других вычислительных подходов к решению задачи состоят в следующем:

1.Используемое представление молекул основано на пространственном *распределении плотности молекулярной формы*, а также от положения потенциальных мест образования водородных связей, т.е. не зависит от связи атомов "внутри" молекулы;

2.Проводится *автоматический выбор* представления данной молекулы *среди многих возможных конформаций* и ориентаций молекулы, что является определяющим фактором при решении задачи;

*3.*Используется *нелинейная модель* при построении зависимостей "структура-активность" - *вычислительная нейронная сеть*.

Для сопоставления и выбора конформаций двух молекул (из обучающей БД) в алгоритме Compass используется *мера близости*, зависящая от взаимного положения поверхностей Вандерваальса двух структур и от расположения функциональных групп - потенциальных доноров и акцепторов водородной связи [32]. Точность численных предсказаний активности по построенной модели рецептора существенно зависит от комбинации трех вышеперечисленных характеристик алгоритма Compass [45]. Весь процесс построения модели (для БД, содержащей около двух десятков структур) занимает около одной недели, при этом основное время компьютера тратится на поиск и анализ *конформационного пространства обучающей БД*, в которой *допустимо присутствие гибких структур*, что существенно выделяет этот метод среди прочих.

Для сокращения времени построения и поиска конформаций молекул обучающей БД, комплементарно соответствующих активному центру рецептора, Willett с соавторами предложил использовать *генетический алгоритм* [46, 47], который реализован в новой программе GASP (Genetic Algorithm Similarity Program)[48]. Назначение программы GASP - пространственное сопоставление конформаций гибких молекул и обнаружение фармакофора на этой основе. Программа демонстрирует высокое качество сопоставления

пространственных форм молекул обучающей БД. С ее помощью были заново построены известные из литературы фармакофоры для ряда рецепторов: 5-HT_{1D}, агонисты дофамина, бензодиазипиновый и энкефалиновый рецепторы [48].

Фармакофорный поиск в БД с трехмерными структурами. Знание структуры активного центра рецептора позволяет проводить скрининг структурных баз данных, содержащих сведения об уже синтезированных молекулах. Примерами программ, успешно использующих такой скрининг для поиска новых лекарств, являются программы DOCK [49] и CLIX [50]. По таблицам связности расчитываются трехмерные структуры, что позволяет идентифицировать молекулы, потенциально обладающие искомой биологической активностью, экспериментальное измерение которой не проводилось. Был разработан ряд методов для проведения такого 3D-поиска [51,52]. Идея использования БД с трехмерными структурами молекул для поиска новых классов структур, обладающих заданной биологической активностью, была предложена в середине 70-х годов[53,54]. Разработка соответствующего программного обеспечения и интенсивные исследования возможностей 3D-поиска для предсказания биоактивности и конструирования лекарств начинается во второй половине 80-х годов [55,56,57,58,59,60,61]. Примерно в это же время создаются алгоритмы и программы молекулярной механики [62,63], позволяющие получить существенный выигрыш во времени вычислений пространственных конформаций по сравнению с квантово-химическими расчетами. Появляются алгоритмы и программы - CONCORD, WIZARD, - для быстрой генерации (в течении нескольких секунд) трехмерной структуры молекулы по ее топологическому представлению в виде таблицы связности атомов [64,65,66,67]. В работе Crippen [68] приведен обзор возникающих здесь задач, включая проблему расчета конформаций макромолекул. В 90-х годах 3D-поиск становится процедурой, которой уделяется все большее внимание [69] в связи со стремительным ростом возможностей вычислительной техники. При этом особое значение приобретает пространственное представление гибких молекул с учетом возникающих неоднозначностей и возможности значительных конформационных изменений лиганда при связывании его с рецепторов (белком) [70,71,72]. Одно из возможных решений проблемы было практически

реализовано в Национальном Онкологическом Институте США, где структурная БД, содержащая более 450 тысяч структур, была преобразована в форму, пригодную для проведения 3D-поиска потенциально активных структур с учетом допустимых множественных конформаций гибких молекул, хранящихся в БД[73]. Расчет множества таких конформаций был проведен с использованием программы Chem-X [61]

1.3 Построение QSAR-моделей - количественных корреляций "структура - биологическая активность"

Поиск зависимостей "структура-свойство" основано на использовании методов распознавания образов[74,75]. Молекулярная структура представлена *вектором признаков*, а обучающая база данных(СБД), содержащей структуры молекул, описывается соответствующей таблицей (матрицей) “молекула-признак” [76,77,78,79,83,145]. Столбцы этой матрицы соответствуют признакам, а строки - молекулам.

При решении каждой конкретной QSAR задачи необходимо найти такие признаки, которые существенно влияют на изменение исследуемого свойства (активности) в анализируемой обучающей СБД. Поиск зависимостей может вестись в различных классах функций: линейных, квадратичных, логических, вероятностных, с помощью нейронных сетей. QSAR модель представляет собой функциональное выражение, вычисляемое на признаках молекул и аппроксимирующее экспериментальные значения химического свойства.

После представлении молекулы *вектором признаков* обучающая БД будет описываться **N**×**M** *таблицей "молекула-признак"* $\mathbf{X} = ||\mathbf{x_{ij}}||$. Строки таблицы соответствуют молекулам обучающей БД (*i=1,...,**N***), а столбцы - признакам, выбранным для описания молекулы (*j=1,...,**M***). Значения свойства/активности образуют вектор столбец $\mathbf{Y}= (y_1,...,y_N)^T$, где y_i - экспериментальное значение свойства i-молекулы обучающей БД. Задача состоит в том, чтобы используя методы анализа данных[80] выразить значения вектора столбца **Y** свойства через вектор-столбцы матрицы **X** с минимальной ошибкой и максимальной прогностической способностью [81]. Таким образом, при характеристике той или иной QSAR модели необходимо указать:

1.Какие признаки были использованы для описания молекулы?

2.В каком классе функций была построена QSAR зависимость?

Анализ связи "структура - биологическая активность" как концепция QSAR моделирования при конструировании лекарств была введена в начале 60-х Hansch с сотрудниками [76,82,83] , а также Free и Wilson [84]. Это статистическое направление в анализе связи "структура-активность". Основное содержание ***метода Ханша*** [85] состоит в том, что эмпирическая модель биологической активности основывается на *линейной зависимости свободной энергии исследуемого процесса* от физико-химических параметров, рассматриваемых как независимые переменные. Метод основан на предположении о существовании корреляции между факторами, определяющими биологическую активность, и физико-химическими параметрами фрагментов-заместителей в ряду родственных химических соединений, содержащих общий структурный фрагмент[86]. Все физико-химические факторы, связанные с транспортными свойствами и взаимодействиями с активным центром, слагаются из трех составляющих: (1) гидрофобной, (2) электронной, (3) стерической. Вклад каждой из этих составляющих характеризуется с помощью соответствующих констант заместителя, описывающих различие в свойствах между первым членом гомологического ряда и рассматриваемым соединением [83]. Электронные свойства описываются константами Гаммета или другими электронными параметрами заместителей. Стерические свойства характеризуются стерическими параметрами заместителей, например, стерической константой Тафта (Es). Гидрофобность соединения описывается логарифмом коэффициента распределения LogP соединения между водой и фазой, моделирующей липид (обычно, нормальным октиловым спиртом). Первоначально применялись линейные модели, позже на практике получили распространение также и полиномиальные зависимости. Цель подхода Ханша - идентификация параметров, характеризующих связь гидрофобных, электронных и стерических факторов заместителей с биологическим свойством молекулы. Эти параметры используются для построения линейных моделей первого и второго порядков, описывающих наблюдаемые изменения активности соединений.

Аддитивная модель Фри-Вильсона основана на предположении, что вклад заместителя, находящегося в молекуле в заданном положении, всегда одинаков, независимо от того, в каком соединении рассматриваемого химического класса присутствует этот заместитель [84]. Тогда биологический отклик соединения может быть представлен как сумма активностей заместителей плюс некоторая средняя общая активность. Величины вкладов заместителей рассчитываются с помощью множественного линейного регрессионного анализа. При построении модели рассматривается базовая молекулярная структура и биологическая активность соединений (физико-химические данные не используются). Было показано, что аддитивная модель Фри-Вильсона и соотношение линейности свободной энергии Ханша теоретически эквивалентны.

Метод Ханша фактически представляет собой физико-химический подход к QSAR-моделированию: локальные физико-химические свойства заместителей (как признаки молекулы) связываются с биологической активностью. В нашей стране этот подход развивается в работах Раевского О.А. с коллегами[87,88,138,180], которые разработали унифицированную шкалу для оценки донорно-акцепторных факторов как локальных свойств молекулярной структуры, а также алгоритмы и программы для их вычисления[89,90,91]. Существует много различных программных систем, ориентированных на использования метода Ханша. Ниже, как пример, описаны две из них - OASIS и EDISFAR.

Система OASIS. Метод Ханша выбран в *программной системе OASIS* (Optimised Approach based on Structural Indices Set) в качестве исходной теоретической предпосылки для описания молекул. Система OASIS [92] создана в Институте Химической Технологии (Бургас, Болгария) и предназначена для предсказания биологических активностей химических веществ (таких, как лекарственная и пестицидная активности, токсичность и т.д.). Для описания структур используются три типа индексов заместителей: электронные (константы Гаммета), стерические (константа Тафта) и показатель гидрофобности. Для описания взаимного влияния заместителей на исследуемое свойство используются локальные “электронные” дескрипторы, включающие в себя заряды и индексы делокализации. Для описания окружения атомов используются физико-химические индексы (константы

Гаммета, фактор гидрофобности, молекулярная рефракция и т.д.) и многомерные стерические индексы Верлупа. Использование (вместе с тремя основными параметрами метода Ханша) большого количества дополнительных индексов является, по мнению авторов, основным достоинством системы. Вместе с тем, следует заметить, что вычисление некоторых используемых в системе характеристик, таких как фактор гидрофобности, молекулярная рефракция, само по себе является самостоятельной и нетривиальной задачей. Для отбора значимых дескрипторов в системе используется стандартный кластер анализ и метод пошаговой регрессии. Система OASIS состоит из 12 программных модулей, обеспечивающих ввод и хранение химических структур и данных, вычисление топологических и стерических индексов, геометрическую оптимизацию молекул, вычисление некоторых квантово-химических (локальных и глобальных) характеристик молекул, кластеризацию обучающей выборки и проведение пошаговой регрессии в каждом отдельном кластере[93].

Система EDISFAR. Программная система предназначена для анализа методом Ханша обучающих СБД, содержащих серии химически родственных соединений. Система EDISFAR затем используется для конструирования новых структур, принадлежащих исследуемому классу[94]. Основой системы является БД фрагментов-заместителей MONOBASE, содержащая для каждого фрагмента *список его локальных физико-химических и расчетных теоретических параметров*. Пользователь системы выбирает важные (с его точки зрения) свойства заместителей, и система проводит кластеризацию фрагментов MONOBASE по выбранным параметрам. Определяются группы заместителей, существенно отличающихся друг от друга в заданной метрике. При конструировании новых молекул с заданным (в обучающей БД) общим заместителем проводится выбор фрагментов из построенных кластеров. Система написана на языке Borland Pascal 7.0 и Turbo-Vision 2.0 на основе объектно ориентированного подхода и предназначена для работы на персональных компьютеров IBM PC [95].

Описание *современного состояния дел* при использовании QSAR подхода Ханша, обзор полученных результатов и соответствующего программного обеспечения можно найти в монографии Kubinyi [96].

1.3.1 Параметрическое описание молекул

Первоначально QSAR параметры классифицировались на три группы: липофильные, электронные и стерические [79], затем эта классификация была расширена [97]- были добавлены теоретико-графовые (formula-derived) параметры (топологические индексы молекулярных графов) и характеристики, влияющие на растворимость.. Таким образом, в общем случае описание структуры представляется в виде вектора параметров { x_i, i=1,...,M}. При этом параметр x_i может быть дискретным (например, вхождение заданного структурного фрагмента) или количественным (например, значение физико-химической характеристики). Можно выделить следующие способы описания химической структуры, базирующиеся на использовании таких признаков , как [98]:

1.константы заместителей и другие физико-химические характеристики отдельных групп;

2.структурные фрагменты - сюда относятся различные типы молекулярных подграфов, описываемых фрагментарными кодами химических структур. Простейшим набором структурных фрагментов является совокупность заместителей в различных положениях молекулы;

3.топологические характеристики матрицы смежности химического графа (матрица связей) и порождаемые на ее основе инварианты графов;

4.квантово-химические характеристики - заряды на атомах, потенциалы ионизации, различные индексы реакционной способности и т.д.;

Выбор способа описания структуры определяется характером решаемой задачи и существующими ограничениями на получение экспериментальных и расчетных данных. Для классификации и получения обобщений был проведен анализ 56 заместителей, характеризуемых 76 различными дескрипторами, описанными в литературе за последние 25 лет [99]. В результате факторного анализа было найдено, что первый фактор описывает *молекулярный объем*, второй фактор представляет *электронные характеристики* заместителя, а третий - способность образовывать *водородные связи*. На этой основе была найдена *интерпретация липофильности* в векторных терминах в следующем виде: ***липофильность = молекулярный объем + полярность.***

1.3.2 QSAR-анализ трехмерных молекул

Построение зависимостей "структура-активность" при использовании пространственного представления молекул обучающей СБД получило название 3D-QSAR (3D-QSAR). Это направление весьма интенсивно развивается в последние годы [100], благодаря появлению доступной мощной вычислительной техники. Продвижение в области разработки новых методов и алгоритмов построения пространственных форм молекул включает подходы, использующие методологию экспертных систем и искусственного интеллекта [101, 102]. Современное состояние этого направления представлено в сборнике [103]. Основные этапы 3D-QSAR моделирования состоят в следующем [100]:

1.Выбрать молекулы в обучающую БД, каждая из которых обладает активностью, экспериментально измеренной для данной биологической системы;

2.Построить пространственные представления этих молекул и провести их "выравнивание" (alignment) согласно заданным правилам выбора ориентаций;

3.Для всех молекул вычислить набор пространственно зависимых признаков;

4.Построить функцию, выражающую зависимость вычисленных признаков и исследуемой биологической активности;

5.Определить устойчивость и предсказательную способность найденной функциональной зависимости.

6.При необходимости модифицировать модель, повторив этапы 1-5.

Наиболее сложным аспектом 3D-QSAR моделирования является выработка правил пространственного выравнивания конформаций молекул обучающей БД. В большинстве используемых подходов используется гипотеза (определяемая на экспертном уровне) о важности для данной активности некоторого набора (обобщенных) функциональных групп, - фармакофоров, что определяет выбор (задание) этих групп в структурах обучающей БД. Если модель рецептора неизвестна, то используется "подход активных аналогов" (Active Analogue Approach) [104] для генерации

гипотетической модели активного центра, который геометрически соответствует взаимодействиям с активными структурами обучающей БД. Весьма обнадеживающие результаты получены при использовании генетического алгоритма для выравнивания конформаций как по скорости вычислений, так и по качеству выравнивания [48, 105]

CoMFA - Comparative Molecular Field Analysis. Наиболее широко используемым методом в рамках 3D-QSAR парадигмы стал метод "сравнительного анализа молекулярных полей" (CoMFA), предложенный Cramer с соавторами в 1988 году [106]. При классическом CoMFA моделировании после выполнения процедуры пространственного выравнивания каждая молекула помещается в трехмерную регулярную решетку. В узлах решетки вычисляются значения стерического (вандерваальсова) и электронного взаимодействия между данной молекулой и "пробным" атомом, имеющим свойства атома углерода в $\mathbf{sp^3}$ состоянии с зарядом "+1". Получаются две трехмерные матрицы, каждая из которых является дискретным представлением стерического и электронного поля исследуемой молекулы. Значения элементов этих матриц являются признаками молекул обучающей БД. В результате формируется очень широкая (M>>N) таблица "молекула-признак", которая анализируется с помощью метода частичных наименьших квадратов (PLS - Partial Least Square) [107,108]. Фактически, PLS строит регрессионное уравнение "структура-свойство" на главных компонентах (латентных переменных) таблицы "молекула-признак"[109,110], а оптимальное число латентных переменных, включаемых в уравнение, определяется на основе процедуры скользящего контроля (cross-validation) [81,111]. Именно эта процедура обеспечивает формирование прогностически устойчивых моделей[112]. Одним из важных достоинств метода является возможность (после возврата к исходным признакам) определения пространственных участков вокруг молекулы, изменения стерического и/или электронного поля в которых приводит к существенным изменениям в биологической активности. Наиболее тонким моментом в использовании CoMFA является необходимость "пространственной нормализации" молекул обучающей БД, т.е. выбор их расположения (после взаимного пространственного выравнивания) относительно системы координат, поэтому до последнего времени метод использовался для анализа жестких и

полужестких молекул. Недавно было обнаружено, что даже изменение положения системы координат (например, ее простой поворот) может привести к падению прогностической способности модели в два раза[113].

Развитие CoMFA подхода состояло в введении новых полей, на основе которых сравнивались молекулы. Так, в CoMFA моделировании широко используется поле, предложенное Goodford [114], поскольку оно лучше описывает возможность образования водородных связей, имеющих ключевое значение при связывании лиганд-рецептор. Вместо атома углерода в $\mathbf{sp^3}$ состоянии с зарядом "**+1**" используют и другие типы "пробных" атомов. Так, Kim с соавторами исследовал "пробы" **H+**, $\mathbf{H_2O}$, $\mathbf{CH_3}$ для предсказания констант ионизации $\mathbf{pK_a}$ имидазолов и имидозолинов при тестировании CoMFA подхода на возможность воспроизведения различных (известных) физико-химических констант заместителей [115]. Были успешно тестированы и другие "пробы" [116], порождающие дополнительные молекулярные поля [117], включая поле, представляющее высшую занятую молекулярную орбиталь структуры [118] и поле "гидрофобных взаимодействий" [119]. Последнее поле было построено на основе: (1) инкрементной системы фрагментов Hansch-Leo [83] для аддитивной оценки липофильности веществ (путем умножения значений инкрементов на площадь поверхности атома, доступную для растворителя) и (2) введения экспонентно убывающей функции дистанции. Аналогичное поле "потенциала липофильности" было введено в работе Gaillard с соавторами [120] на основе липофильной системы атомов Broto-Moreau [121,122]. Было предложено использовать CoMFA подход для анализа трехмерных структурных баз данных[123].

Интересно отметить, что проведенное сравнение физико-химического подхода Ханша и CoMFA метода показало, что оба метода показывают примерно одинаковую прогностическую способность[124], однако CoMFA алгоритм позволяет получить картину пространственного распределения стерических и электронных областей, изменения в которых приводят к колебаниям в биоактивности молекул.

Популярность CoMFA моделирования[125,126] вызвана тем, что реализующая его программа включена в коммерчески доступную для различных вычислительных платформ систему молекулярного моделирования SYBYL

фирмы Tripos [37]. В середине 1996 года эта фирма объявила о выпуске нового CoMFA модуля (Advanced CoMFA), усовершенствованного с учетом опыта, накопленного пользователями при эксплуатации SYBYL системы [127]. Выпущенная программа включает в себя следующие новые возможности:

1. Средства анализа. Позволяют вводить весовые коэффициенты на элементы трехмерной решетки и тем самым (1) подчеркивать значимость определенных пространственных участков вокруг молекулы, выявленных в результате первичного PLS-анализа матрицы "молекула-признак"[128], а также (2) маскировать "шумовые" участки. Ускорено выполнение процедуры скользящего контроля, что существенно повышает эффективность оптимизации модели.

2. Введены средства текущего оперативного протоколирования этапов CoMFA моделирования в виде электронных таблиц (spreadsheet), в которых автоматически заносятся результаты PLS, SIMCA [129], иерархического кластерного анализа [130] и QSAR анализа данных.

3. Средства пространственного выравнивания. Автоматически проводится выравнивание всех молекул обучающей БД к заданной структуре-шаблону.

4. На основе построенной 3D-QSAR модели автоматически генерируются и автоматически пространственно оптимизируются новые молекулы, удовлетворяющие заданной модели, т.е. для которых прогноз исследуемой биологической активности имеет высокое значение (проводится решение обратной QSAR задачи).

5. Включены новые классы молекулярных полей, включая поля H-связей (донорные и акцепторные), индикаторные и параболические поля.

6. Включены средства проведения иерархического кластер анализа с возможностью гибкого формирования групп на уровне обобщения, задаваемом пользователем, что позволяет оперативно определять области вокруг молекулы, в которых исследуемое поле обладает заданными локальными свойствами.

1.3.3 Трехмерные дескрипторы молекул

Разработка признаков, способных содержательно характеризовать пространственную структуру биологически активных молекул, началась еще в

конце 70-х годов, в рамках использования общей методологии распознавания образов[131,132] в применении к химии [133]. Одна из ранних попыток охарактеризовать пространственную геометрию молекулярных структур обучающей БД была предпринята Soltzberg с соавтором в 1977 году и состояла в введении так называемого "молекулярного преобразования" по формуле:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} Z_i Z_j \frac{\sin(s r_{ij})}{s r_{ij}}$$

где N - число атомов в молекуле, включая атомы водорода; r_{ij} - межатомное геометрическое расстояние; Z_i - порядковый номер элемента в периодической таблице для i-ого атома; s - параметр рассеивания. Функция $I(s)$ введена по аналогии с формулой расчета интенсивности рассеивания молекулой ренгеновского излучения [134].

Функция "молекулярного преобразования" $I(s)$ была использована для формирования двоичного вектора следующим образом: интервал s (0<s<31) был разбит на 100 равных участков, если на участке $I(s)$ (как функция s) обращалась в ноль, то булевской переменной, соответствующей этому участку, присваивалось значение "1". Таким образом, каждой структуре обучающей СБД, состоящей из 219 соединений, проявляющих седативный (79 молекул) или транквилизирующий (140 молекул) эффект, был поставлен в соответствие двоичный вектор, содержащий 100 компонент. Построенная дискриминантная функция на скользящем контроле (10 прогонов по схеме "leave-out-ten") показала 90% правильных предсказаний, что являлось весьма хорошим результатом для такой большой выборки весьма химически разнородных соединений. Эта выборка ранее исследовалась Jurs с соавторами [135], которые получили QSAR-модели с сопоставимым прогностическим качеством. QSAR исследование молекулярного преобразования для топологической матрицы расстояний в 90-х годах проводил King с соавторами [136,137].

Основываясь на функции молекулярного преобразования $I(s)$, Новиков и Раевский в 1982 году ввели функцию радиального распределения $f(r)$ [138]:

$$f(r) = \frac{2}{\pi} \int_0^{s_{max}} sI(s) \exp(-bs^2) \sin(sr) ds$$

Эта функция представляет собой спектр межатомных расстояний, интенсивность которого определяется видом атомов, находящихся на данном расстоянии друг от друга, и динамикой изменения этого расстояния в процессе внутримолекулярных колебаний. Современная вычислительная техника позволяет расчитывать на ПЭВМ и использовать (как QSAR признаки) функции ***I(s)*** и ***f(r)***, достаточно сложные в вычислительном отношении. Представляется перспективным систематическое исследование их поведения на обучающих БД различных химических классов.

MSA - Molecular Shape Analysis В 1980 году Hopfinger предложил проводить QSAR исследование биологически активных соединений с помощью анализа молекулярных форм (MSA) [139]. Для каждой пары (i,j) молекул обучающей БД (после попарного выравнивания конформаций) вычислялась величина $\mathbf{V_0(i,j)}$ взаимного пересечения молекулярных объемов по формуле:

$$V_0(i,\ j) = \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (v_{i,k} \cap v_{j,l})$$

где n_i, n_j - число атомов в **i**-ой и **j**-ой молекулах, $v_{i,k}$ - вандерваальсов объем **k**-ого атома в **i**-ой молекуле. На основе параметра $\mathbf{V_0(i,j)}$, имеющего размерность объема, для QSAR анализа вводились признаки $\mathbf{S_0(i,j)}$ и $\mathbf{L_0(i,j)}$, имеющие размерность площади и длины соответственно:

$$S_0(i,\ j) = [V_0(i,\ j)]^{2/3};\ L_0 = [V_0(i,\ j)]^{1/3}$$

Для формирования описания обучающей БД одна из структур выбиралась в качестве опорной (reference structure), и относительно ее расчитывались величины $\mathbf{V_0}$, $\mathbf{S_0}$, $\mathbf{L_0}$ для всех остальных молекул как параметры таблицы "молекула-признак". Получаемые с помощью этого подхода QSAR модели по прогностической силе не уступают CoMFA методу [140]. В последнее время MSA метод является одним из подходов к вычислению пространственного молекулярного подобия при решении QSAR задач [141].

Для характеристики пространственных особенностей фрагментов-заместителей используются экспериментальные значения вандерваальсовых радиусов атомов[142]. Verloop разработал стерические дескрипторы [143], которые широко используются и хорошо себя зарекомендовали в QSAR анализе серий родственных структур методом Ханша[96].

Для описания пространственных межмолекулярных взаимодействий, которые являются полярными по своей природе и вносят существенный вклад в экспериментально наблюдаемые свойства, Stanton и Jurs предложили использовать серию трехмерных структурных параметров, в которых объединена информация как о молекулярной поверхности, так и частичных зарядах на атомах - CPSA-дескрипторы (Charge Partial Surface Area)[144]. Для каждого атома после оптимизации геометрии молекулы (методом молекулярной механики) вычисляются два параметра (площадь поверхности и заряд): SA_i и Q_i , которые разделяются на положительные ($+SA_i$ и $+Q_i$) и отрицательные ($-SA_i$ и $-Q_i$). На их основе определяются следующие параметры (Partial Positive/Negative Surface Area):

$PPSA\text{-}1=\Sigma(+SA_i)$ $PNSA\text{-}1=\Sigma(-SA_i)$

$PPSA\text{-}2=[\Sigma(+SA_i)]^x (+Q_T)$ $PNSA\text{-}2=[\Sigma(-SA_i)]^x (-Q_T)$

$PPSA\text{-}3=\Sigma[(+SA_i)^x (+Q_T]$ $PNSA\text{-}3=\Sigma[(-SA_i)^x (-Q_T)]$

где $+Q_T=\Sigma(+Q_i)$, $-Q_T=\Sigma(-Q_i)$,

Дескрипторы разности положительной и отрицательной площади молекулярной поверхности вычислялись по формулам:

DPSA-1=(PPSA-1)-(PNSA-1); DPSA-2=(PPSA-2)-(PNSA-2);

DPSA-3=(PPSA-3)-(PNSA-3)

Модуль вычисления этих дескрипторов был введен в QSAR систему ADAPT[145, 146]. Система ADAPT имеет в своем составе многочисленные молекулярные дескрипторы разных типов (топологические, подструктурные, геометрические)[145] и хорошо себя зарекомендовала для прогнозирования биологической активности (психотропная, запах и др.)[147].

Описание других методов формирование пространственных дескрипторов молекул для QSAR анализа можно найти в работах [100,103].

1.3.4 Подструктурные молекулярные дескрипторы.

Большая группа подходов к описанию веществ в виде вектора признаков связана с фрагментацией молекулярной структуры, т.е. с разложением ее на некоторые фрагменты-дескрипторы, выбираемые из условия простоты выделения или содержательных соображений [148]. Методы фрагментации

первоначально рассматривались в химической документалистике применительно к задаче информационного поиска[149,150,151]. Затем они были перенесены на задачи анализа QSAR методами статистической теории и теории распознавания образов.

Подструктурный анализ [152] базируется на предположении, что биологическое действие вещества обусловлено наличием в его составе некоторых элементов строения (подструктур). Структура соединения представляется фрагментарным кодом [149]. Относительный вклад подструктур в активность определяется отношением числа активных веществ, содержащих данную подструктуру, к общему числу соединений в исследуемой серии. Активность соединения определяется как среднее арифметическое вкладов подструктур, составляющих тестируемое соединение. Среди используемых типов фрагментов можно отметить следующие:

1.атомы и пары связанных атомов [135,153];

2.присоединенные атомные фрагменты, имеющие центральный атом, характеризуемый его связями и присоединенными к нему атомами [154,155,156]. К этому типу фрагментов относятся “ганглиофрагменты”, дополнительно включающие в себя соседей второго уровня [157];

3.циклические фрагменты, характеризующие форму циклов .и положение в них гетероатомов [156,147,158, 159];

4.фрагменты "гетеропутей", описывающие в молекуле цепи, начинающиеся и кончающиеся гетероатомами [156];

5.фрагменты в линейной нотации Висвессера [160, 161, 149].

6. фрагменты расширенного языка "брутто-формул связей" [151,162] с введением модификаций микрофрагментов[163];

7.подструктуры, характерные для исследуемой СБД [164,165],166];

Методом подструктурного анализа была исследована обучающая СБД, содержащая 850 структурно разнородных соединений, испытанных на физиологическую активность[158]. Были выделены наиболее информативные подструктуры, которые могут быть использованы для прогнозирования активности неиспытанных соединений. Анализ подструктурного подхода для описания химических объектов изложен в работе[167]. Использование подструктур является типичным примером подхода к решению сложных

проблем, зависящих от представления объекта. Существуют различные способы классификации фрагментов, когда каждый тип подструктуры идентифицирует определенный аспект молекулы, а использование того или иного типа фрагментов зависит от характера решаемой QSAR задачи. В разрабатываемой с середины 70-х годов[168] системе DARC для подструктурного кодирования молекул используется порождающая концепция формирование фрагментов FREL (Fragment Reduced to an Environment that is Limited). [169]. Она лежит в основе унифицированной модели формирования фрагментов и служит для дальнейшего развития подструктурного представления и классификации молекул.[170]

Фрагментарный код суперпозиции подструктур. Для решения задач QSAR анализа был разработан специальный язык *фрагментарного кодирования* химических структур - фрагментарный код суперпозиции подструктур (ФКСП) [171, 172, 173]. Выделяемые в этом языке фрагменты характеризуют циклы, пути между гетероатомами и положение заместителей при ароматических и гетероароматических системах. При построении ФКСП было использовано понятие "топологического фармакофора", которое базируется на следующих положениях[98]:

1.Свойство химических соединений оказывать фармакологическое действие определяется способностью комплементарно фиксироваться и взаимодействовать с функциональной химической структурой в живом организме "рецептором".

2.Во взаимодействии химических соединений с рецептором принимает участие не вся молекула, а какая-то ее часть "фармакофор". Это взаимодействие ограничивается образованием относительно слабых химических связей: водородных, электростатических, вандерваальсовых, гидрофобных, ионных.

3.Соединения, обладающие одинаковыми или сходными фармакофорами, способны оказывать одинаковое специфическое действие.

Был сформирован список из 37 потенциально активных центров ("дескрипторных центров" - ДЦ), в который вошли атомы или группы атомов, способные участвовать в "слабых" взаимодействиях. Дескрипторными центры ФКСП являются все гетероатомы: **N, O, S, P**, галогены, металлы,

ароматические циклические системы в целом, а также пары атомов углерода, связанные кратными связями (двойными и тройными, но не ароматическими). *Дескриптором ФКСП является цепочка атомов углерода, начинающиеся и кончающиеся ДЦ.* ФКСП-код строится по схеме:

ДЦ1	***Длина цепи (в атомах углерода)***	***ДЦ2***	***Признак сопряжения***

ФКСП-код представляется 7-значным целым числом: по две цифры отводится под ДЦ и под длину цепи, одна цифра кодирует признак сопряжения (1 - наличие, 0 - отсутствие сопряжения). Циклические системы кодируются в языке ФКСП с помощью *специальных циклических дескрипторов* в следующем виде:

"Голова"	**"Ядро"**	**"Хвост"**
Геометрическая форма циклической системы	*Кол-во пи-электронов в циклической .системе*	*Расположение гетероатомов*

"Голова" описывает размер цикла (число атомов в случае моноцикла, или размеры и расположение отдельных простых циклов в случае полициклической системы). "Хвост" обозначает гетероатомы и их положение в кольце, которое задается с помощью канонической нумерации атомов в циклической системе. Был введен специальный список циклических систем (мостиковые, трехмерные и т.п.), которые не могут быть закодированы по перечисленным выше правилам, и поэтому в качестве кода указывается их номер по списку.

Развитием языка ФКСП явилось введение *матрицы связности дескрипторных центров* (МСДЦ). Структуру соединения можно описать в виде неориентированного помеченного графа, вершинами которого являются ДЦ. Метками вершин являются коды ДЦ, а метками ребер - длины путей между ДЦ. Такой граф *называют графом связности дескрипторных центров*, а его матрицу смежности, содержащую метки вершин и ребер МСДЦ-матрицей. В *МСДЦ отсутствуют специальные дескрипторы для циклических систем*, поэтому для полноты учета информации о циклах считают, что ДЦ есть любой единичный цикл, содержащий не более 9 атомов.

ФКСП дескрипторы используются в логико-структурном методе конструирования лекарств. *Логико-структурный подход* (ЛСП) основан на миллевской логике [174,175] и позволяет определять стохастические закономерности[176]. Для ЛСП характерно моделирование некоторых сторон мышления исследователя при установлении связи между структурой и активностью[177]. Алгоритмы ЛСП позволяют автоматизировать анализ QSAR и осуществлять его корректно[178]. ЛСП основывается на следующих допущениях:

•Химическая структура может быть представлена с помощью некоторого формального описания.

•Соединения могут быть разделены в соответствии с проявленной ими активностью на N классов. В общем случае классы активности могут пересекаться, т.е. некоторые соединения могут одновременно принадлежать к различным классам активности.

•Признаки биологической активности определяются как ФКСП-фрагменты, присутствие которых с достаточно высокой вероятностью указывает на то, что исследованное соединение проявляет данную активность.

Алгоритм отбора ФКСП-признаков имеет следующую структуру[179]:

1.Потенциальные признаки активности отбираются путем нахождения общих ФКСП-фрагментов для каждого класса активности.

2.Подсчитывается общее число появлений ФКСП-признака в каждом из классов.

3.Проводится статистическая оценка ФКСП-признаков для каждого класса на основе общего числа появлений признака в обучающей БД.

4.Признак считается разделяющим, если принятые для него статистические оценки превышают некоторый порог. Эти пороги могут оптимизироваться с применением процедуры "скользящего контроля".

ФКСП-кодирование молекул и логико-структурный подход были положены в основу программной системы ОРАКУЛ (Оптимизированный Распознающий Алгоритм Конструирования Усовершенствованных Лекарств), предназначенной для конструирования лекарств по большой БД. База данных содержала несколько тысяч структур, описанных в виде ФКСП. Распознавание активности в системе проводится по ФКСП-фрагментам, статистические

характеристики которых превосходят заданный порог. Развитием системы ОРАКУЛ явилась система ТОПЛОГ. В отличие от системы ОРАКУЛ, новая система использовала несколько языков описания структуры: язык команд для интерактивного ввода структурных формул, компактную форму таблицы связей для хранения структуры, код ФКСП-2 для быстрого поиска подструктур, топологической представление матрицы связей дескрипторных центров для отбора признаков биологической активности [98].

Раевский с соавторами предложил детализировать классификацию ДЦ, используемых для построения ФКСП-кода, на основе использование *унифицированной шкалы донорно-акцепторных взаимодействий* [180].

Понятие топологического фармакофора как подструктуры специального вида интенсивно развивалось для прогнозирования биологической активности в работах Franke и его сотрудников, которые разработали методы их компьютерного анализа LOGANA и LOCON[181, 182]

Таким образом, разбиение молекул на фрагменты на основе "фармакофоров" сводится к определению *заранее фиксированного списка* дескрипторных центров. При анализе специфической биологической активности (например, противоопухолевых препаратов), этот список необходимо будет *существенно откорректировать*: например, различать все галогены. При переходе к анализу антивирусных препаратов может понадобится третья классификация фармакофоров и т.д. В связи с этим при разработке программного обеспечения, предназначенного для исследования биологической активности разной природы, необходимо предусмотреть возможность динамического введения той или иной классификации фрагментов как дескрипторных центров.

Были предложены и успешно использованы для поиска в БД биологически активных молекул подструктурные дескрипторы “атомные пары” (AP-Atom Pairs) и “топологическая скрученность”(TT - Topological Torsions) [183, 184, 185]. *AP-подструктуры* имеют вид:

(atom type i)-(separation)-(atom type j)

где ***(atom type i)*** содержит информацию: (1) об *имени атома* (по периодической системе), (2) о числе его *смежных неводородных атомах* и

(3) о числе его *пи-электронов*; ***(separation)*** есть *топологическое расстояние* между атомами **i** и **j**. Молекула, имеющая **N** неводородных атомов, будет представлена множеством из **N(N-1)/2** АР-подструктур. *ТТ-подструктуры* состоят из четырех атомов **i**, **j**, **k**, и **l,** которые последовательно связаны в молекуле, и имеют вид:

(atom type i)-(atom type j)-(atom type k)- (atom type l)

Молекула, содержащая **M** наборов последовательно соединенных четырех атомов, будет описана множеством из **M** ТТ-подструктур. АР- and ТТ-фрагменты представляют *комплементарные аспекты* химической структуры. Так, АР-фрагменты учитывают "дальнодействующие" особенности молекулы, в то время как and ТТ-фрагменты описывают ее локальные подструктурные особенности. Важным обстоятельством является то, что АР- и ТТ-дескрипторы формируются автоматически, и для их построения не требуется указаний эксперта, решающего какие подструктуры следует отнести к заместителям, а какие к "молекулярному ядру".

АР- и ТТ-дескрипторы использовались для поиска в структурных БД молекул, сходных с заданной [183]. *Близость* определялась как общее число дескрипторов (АР или ТТ), имеющееся у двух молекул, и после нормирования принимала значения в интервале от 1.0 (идентичные структуры) до 0.0 (нет общих подструктур). Кроме этого, для скрининга больших СБД было введено понятие *"тренд вектора"* [184]. Его компоненты вычислялись на основе имеющейся обучающей БД. Его **k**-ая компонента имеет тем большее *положительное* значение, чем более **k**-ый дескриптор (АР или ТТ) ассоциирован с *активными* структурами обучающей БД. Отрицательные значения векторной компоненты тренд вектора соответствуют ее тесной связи с неактивными молекулами обучающей БД. *Проекция векторного описания молекулы на тренд вектор позволяет ранжировать все структуру СБД,* т.е. дать предварительную оценку вероятности, что структура будет обладать исследуемой активностью. Очень важной особенностью использования тренд вектора является неявное определение области применимости оценки свойства на его основе. Так, необходимо, чтобы не менее 95% дескрипторов

оцениваемого (проецируемого) вектора присутствовало в тренд векторе. При нарушении этого условия оценка активности молекулы не проводится [185].

Для автоматического анализа биологических или физико-химических свойств молекул обучающей БД и *поиска структурных дескрипторов, наилучшим образом описывающих заданное свойство* предназначена программная система Multi-CASE (Multiple Computer Automated Structure Evaluation program) [186], разработанная на Химическом факультете Кливлендского университета (Огайо, США). Структуры обучающей выборки *автоматически разбиваются на все возможные линейные структурные фрагменты* заданной сложности, которые кодируются путем перечисления атомов, составляющих фрагмент, и связей между ними. В качестве значения дескриптора используется количество фрагментов определенного вида в молекуле. Для построения ККСС-уравнения используется стандартный дискриминантный анализ и линейная регрессия.

Основным преимуществом системы Multi-CASE является возможность работы с обучающими выборками, составленными из химических структур *различных классов*. При этом система производит автоматическую кластеризацию выборки. Благодаря этому система применялась авторами для анализа выборок, состоящих из нескольких тысяч соединений. Данный подход к описанию молекул на основе автоматической генерации структурных фрагментов показал на ряде свойств хорошую прогностическую устойчивость.

Кроме вышеописанных типов признаков при построении QSAR зависимостей широко используются топологические индексы (ТИ) - числовые характеристики, являющиеся инвариантами молекулярных графов [187]. При этом делается предположение, что объяснение поведения такой сложной системы как молекула может быть описано в терминах поведения ее более простых компонент (свойства подструктур и их взаимоотношений в молекуле). Описание различных типов топологических, применяемых в QSAR исследованиях описано в следующем разделе.

1.4 Построение количественных корреляций "структура- физико-химическое свойство"

1.4.1 Аддитивные подходы оценки физико-химических свойств веществ

Методы массового расчета свойств многих десятков тысяч молекул и веществ в настоящее время могут быть осуществлены только как *полуэмпирические*, т. е. требующие использования экспериментальных данных по соответствующему свойству в рассматриваемых рядах молекул для определения параметров в ККСС формулах. [188]

Фрагменты молекул с успехом используется для нахождения многих физико-химических свойств органических соединений по аддитивным схемам [188, 189, 190, 191, 192, 193]. "Давно известно, что большинство свойств сравнительно сложных молекул можно представить в виде суммы соответствующих вкладов индивидуальных атомов и связей. Подобные эмпирические закономерности основаны на том, что силы, действующие между атомами внутри молекулы или между атомами различных молекул, являются очень близкодействующими. Из-за этого индивидуальные атомы в любой молекуле дают почти постоянные вклады в такие свойства молекул как коэффициент преломления, ультрафиолетовые и инфракрасные спектры поглощения, магнитная восприимчивость, а также энтропия, молярная теплоемкость и даже теплота образования" [191]."Существует много экспериментальных данных, показывающих, что в молекуле во многих случаях парциальные свойства, которые можно приписать отдельным структурным элементам, в большой степени остаются постоянными от молекулы к молекуле некоторого ряда, т.е. соответствующие структурные элементы по некоторым своим парциальным свойствам очень близки в разных молекулах ряда. Примером этого является большое постоянство парциальных величин энергии, энтальпии, рефракции и некоторых других величин для группы "-CH2-" в молекулах алканов и многих других рядах органических соединений" [194].

Разбиение молекулы на фрагменты (структурные элементы) для описания органических соединений основывается на втором постулате классической теории химического строения [194], который состоит в том, что отдельные структурные элементы, входящие в разные молекулы могут быть классифицированы так, что на структурный элемент определенного вида

приходится приближенно *одинаковая парциальная величина* рассматриваемого свойства, которое имеет характер *фактора емкости*. Все группы структурных элементов в разных молекулах, относимые при определенной классификации к одному виду, принимаются приближенно *эквивалентными*, и каждой такой группе приписывается *одинаковое парциальное значение* рассматриваемого свойства, характеризующего молекулу.

При классификации атомов в молекуле можно учитывать только *тип атома* (химическую индивидуальность, валентность, распределение сродства по связям), или *его вид* (тип и характеристики первого окружения), или его *оптическую разновидность*, если она может иметь место. Чем детальнее классификация, тем точнее эквивалентность структурных элементов. Таким образом, *можно рассматривать разные приближения в возможных классификациях структурных элементов* и разные приближения для степени их эквивалентности."

Выбор критерия фрагментации. Одним из ключевых вопросов аддитивного подхода является выбор способа "дезагрегирования" структуры, поскольку необходимо учитывать двойственность требований к степени фрагментации. С одной стороны, степень разбиения на фрагменты должна быть максимальной, чтобы была возможность учесть такие факторы, как тонкие различия в строении молекул. Такой уровень дезагрегирования приводит к *огромному числу* признаков-дескрипторов и определенным сложностям реализации программы на ЭВМ. С другой стороны, структура должна быть представлена относительно крупными фрагментами. Однако, тогда появляется вероятность того, что будут не выявлены общие признаки активности структуры, поскольку каждая структура индивидуальна. Таким образом, при выборе степени детальности фрагментации *необходимо каждый раз ориентироваться на конкретную задачу*: при рассмотрении различных свойств молекулы *принципы классификации структурных элементов* и групп структурных элементов и *критерии, полагаемые в основу их эквивалентности*, могут быть различными [194].

В монографии Бенсона[191] приведено описание аддитивных схем для расчета различных свойств, разработанных к середине 70-х годов, с обширной библиографией . В этой книге также содержатся подробные таблицы фрагментов с соответствующими парциальными вкладами, которые широко

использовались для проведения вычислений "в ручную", без использования ЭВМ. Большое число аддитивных схем расчета термодинамических свойств газов и жидкостей с соответствующими формулами и таблицами фрагментов приведены в справочнике Рида, Праусница и Шервуда [195]. Описанные методы и расчетные формулы (включающие в себя не только аддитивные схемы, но и расчеты свойств относительно друг друга) ориентированы на вычисления "в ручную", что специально оговорено в предисловии. Так, например, метод расчета на основе топологического индекса связности Рандича [**Ошибка! Закладка не определена.**,196], который обеспечивает высокую точность предсказания для ряда свойств, охарактеризован как вычислительно сложный.

Прекрасным справочником по вычислительным, в том числе и аддитивным схемам расчета экологически важных свойств органических веществ является монография Lyman, Reehl, Rosenblatt [197].

Вопросам построения аддитивных схем для различных классов органических соединений посвящены работы Папулова и Смолякова с соавторами [198,199,200,201]. Ими же получены результаты [202], устанавливающие связь между классическими аддитивными схемами и теоретико-графовыми (топологическими) индексами [203,204,205], которые с середины 70-х годов стали применяться для расчета физико-химических свойств веществ [196].

Современное состояние методов прогнозирования физико-химических соединений превосходно изложено в монографии Horvath [206], содержащей более 1400 ссылок. Horvath описывает существующие компьютерных БД по свойствам органических соединений, а также перечисляет известные программы для расчета различных, важных для инженеров-химиков свойств веществ.

1.4.2 Теоретико-графовые индексы молекулярных структур

В 1878 году Sylvester предложил рассматривать структурную формулу молекулы как граф специального вида [207]. В настоящее время теория графов[208,209,210] успешно применяется в химических исследованиях в таких областях как химическая информатика, различение изомеров, оценка разветвленности молекул, перечисление изомеров, имеющих одинаковую бруто-формулу, вычисление кванто-химических параметров молекул, построение корреляций "структура-физико-химическое свойство", корреляций "структура - биологическая активность" [211,212,213,214,215,216]. Математическое описание химической структуры может быть сделано на основе матрицы, набора чисел или единственного числа. Так, матрица связности A(G), матрица расстояний D(G) и инцидентная матрица T(G) однозначно определяют топологию молекулы. Недостаток матричного представления молекул состоит в том, что матрицу затруднительно использовать для предсказания или коррелирования свойств.

Числовые характеристики, полученные из топологических характеристик молекулярных графов (М-графов) называются топологическими индексами[217]. Важность топологических индексов впервые была установлена в середине 60-х годов, когда они стали применяться в большом числе корреляций и для прогнозирования физико-химических параметров углеводородов и других органических соединений[218].

Первая попытка однозначно охарактеризовать молекулярную топологию одним числом - инвариантом графа, - была предпринята Spialter [219,220] в середине 60-х годов. Он выдвинул гипотезу, что характеристический полином матрицу связности A(G) однозначно представляет граф. Эта гипотеза была опровергнута рядом контрпримеров [221], а такие графы стали называть изоспектальными графами [213]. В настоящее время в исследованиях химических теоретико-графовых индексов наблюдается три различные тенденции [211]:

- одновременное использование нескольких индексов (так называемых супериндексов) с целью адекватной параметризации химической структуры [222];
- разработка индексов с высокой дискриминирующей способностью [**Ошибка! Закладка не определена.**,223 , 224];

- Извлечение содержательной химической информации с помощью известных индексов [225, 226]

В настоящее время дискриминирующая способность индексов для изоморфных структур значительно выросла, и сегодня можно с уверенностью сказать, что они дают фактическое однозначное описание большинства углеводородов. Показано, что топологические индексы отражают в различной степени форму и размер описываемых ими структур. Недостатком представления молекул с помощью топологических индексов является то, что при этом теряются все стереохимические особенности молекулярных структур. Однако известно, что многие важные характеристики молекул, такие как энергия, плотность заряда и другие существенно зависят от топологии молекул.

Относительно природы и использования топологических индексов (ТИ) остаются открытыми некоторые важные вопросы:

1. Поскольку многие известные ТИ взаимно коррелируют друг с другом, то можно ли выделить полезную информацию из ТИ для практического использования в QSAR моделях?
2. Можем ли мы использовать ТИ для упорядочивания молекул и предсказания свойств структурно разнообразных веществ?

Общий взгляд на природу ТИ, их использование в органической химии и способы вычисления ТИ дан в обзоре [211]. Основой построения ТИ являются инвариантные множества М-графов. Каждому такому множеству соответствует целое семейство ТИ. Проведена классификация ТИ [211]:

1.Индексы, связанные с матрицей расстояний (индексы Винера, Балабана);

2.Индексы, основанные на инвариантном множестве степеней вершин (индексы Рандича, Гутмена, Платта, Кира, Кира-Холла (простой и валентный));

3.Индексы, зависящие от спектральных характеристик графа (индекс Ловаша-Пеликана, Хосойи);

4.Информационные топологические индексы.

Один из вариантов классификации структурных и численных инвариантов М-графов приведен в работе [227], где дано обобщение определения

инварианта графа и показана ключевая роль структурных инвариантов - относительных разбиений графа[228], - при построении численных ТИ М-графов. В этой же работе был предложен метод вычисления *относительных ТИ фрагментов*, отличающийся от схемы, введенной в работе Mekenyan et al [229].

Balaban предложил проводить классификацию ТИ на основе введенного им понятия локального атомного инварианта (LOVI - local vertex invariant) [230,231]. Целочисленные ТИ, основанные на целочисленных LOVI, он относит к индексам первого поколения, отмечая что они характеризуются высокой вырождаемостью. ТИ второго поколения являются действительными числами, но вычисляются на основе целочисленных LOVI с помощью специальных функциональных преобразований, как, например, индексы связности Рандича, Кира-Холла и индекс Балабана. Эти ТИ имеют более высокую дискриминирующую способности по сравнению с ТИ первого поколения. Были предложена методология построения ТИ третьего поколения, которые основаны на LOVI, являющихся действительными числами[232,233].

Для дальнейшего развития методов конструирования новых ТИ очень важным является теоретический результат о выражении инварианта произвольного помеченного графа через базис инвариантов, полученный Баскиным, Скворцовой, Станкевичем и Зефировым [234]:

Теорема. Любой инвариант *f(H)* графа H с символьными пометками вершин и ребер однозначно представляется в виде $f(H) = \sum_{j=1}^{N} c_j g_j(H)$, где $g_j(H)$ - число вхождений графа H_j в граф H, $\boldsymbol{c_j}$ - некоторые константы, зависящие только от ***f***.

Эта теорема позволяет при конструировании программной системы вычисления признаков молекул использовать вышеприведенную формулу как основу поиска различных типов новых ТИ.

Индексы матрицы расстояний. Элемент матрицы расстояний d_{ij} для связного М-графа определяется как число ребер в кратчайшей цепи между *i*-м и *j*-м атомами и d_{ii} =0. Тогда индекс Винера[235,248] вычисляется по

формуле $W(G) = \frac{1}{2}\sum_{ij} d_{ij}$. В своей работе 1947 года Wiener определил W как "число путей" (path number), т.е. как сумму дистанций между любыми С-атомами в молекуле алкана в терминах С-С связей [235]. Полученное им уравнение для оценки температуры кипения t_b алканов имело вид:

$$(t_0-t_b)= 98(W_0-W)/n^2+5.5(p_0-p),$$

где **n** - число атомов углерода в молекуле, **p**- "полярное число", т.е. число С-атомов, разделенных тремя связями, W_0, p_0 - значения индексов **W** и **p** для линейного изомера алкана, состоящего из n атомов. Это уравнение имеет две важных особенности. Во-первых, индекс **W** нормирован так, что имеет линейную размерность по числу атомов. К сожалению, в настоящее время проводить такую нормировку, как правило, не принято. Во-вторых, это уравнение фактически представляет серию уравнений для каждой совокупности изомеров с одинаковым числом атомов: при переходе к новой совокупности перерасчитываются "средние" W_0, p_0 .

Анализ показывает, что в случае ациклических систем индекс Винера можно использовать для характеристики степени ветвления и степени компактности графа, а в случае циклических систем степени цикличности. Индекс Винера нашел широкое применение для построения корреляционных ККСС-соотношений, например, при оценках термодинамических характеристик насыщеных углеводородов, в газохроматографическом анализе, при расчете физико-химических свойств полимеров [213,236,237]. Предлагаются различные варианты обобщения и расширения ТИ Винера [238, 239, 240,241], обзор которых приведен в работе [242].

Индекс Смоленского. В 1964 году Смоленским Е.А. была предложена формула для расчета теплоты образования насыщенных углеводородов, основанная на оценке числа цепочек различной длины в М-графе[243]: $f(G) = A_0 + \sum A_k * [X_k]$, где A_k коэффициенты, определяемые методом наименьших квадратов; $[X_k]$ число различных цепочек длины k; $f(G)$ величина физико-химического свойства структуры, представленной графом G, E - ошибка приближения. Подход Е.А.Смоленского явился обобщением классической аддитивной схемы расчета физико-химических свойств углеводородов. Особенностью предложенного подхода явились правила различения цепочек

углеродных атомов. Смоленский ввел дифференциацию цепочек длины три (и четыре) по степеням "внутренних" атомов, что явилось развитием числа полярности Wiener [235].

Матрица *D* позволяет вычислить графа *G*, которые относятся к *метрическим индексам*[244,245,246]. Пусть, что М-граф *G* задан множеством вершин *V* и множеством ребер *E*: *G*=(*V*,*E*), где $|V| = n$ число атомов; $|E| = m$ число связей.

Эксцентриситетные характеристики М-графа:

Эксцентриситет *i*-вершины:	$e_i=max(d_{ij})$;
Радиус графа *G*:	$R(G)=min(e_i)$;
Диаметр графа *G*:	$Diam(G)=max(e_i)$;
Эксцентриситет графа *G*:	$Ex(G)=\sum_{i=1}^{n} e_i$;

Дистанционные характеристики М-графа:

Дистанция *i*-вершины:	$d_i=\sum_{j=1}^{n} d_{ij}$;
Дистанция графа (индекс Винера):	$D(G)=\frac{1}{2}\sum_{i=1}^{n} d_i$;
Средняя дистанция графа:	$Dm(G)=\frac{1}{n} D(G)$;
Индекс Балабана[224,247]:	$J(G)=\frac{2}{\mu+1}\sum_{i,j=1}^{n} 1/\sqrt{d_i \cdot d_j}$, где μ цикломатическое число графа G: $\mu=m-n-1$.

Индексы матрицы смежности. Элемент матрицы смежности *A*(*i*,*j*) равен единице, если ребро (v_i,v_j) присутствует в графе, и нулю в противном случае. Сумма элементов по *i*-строке матрицы смежности есть степень *i*-ого атома ($deg_i=\sum_{j=1}^{n} a_{ij}$). Тогда *индекс связности Рандича* [**Ошибка! Закладка не определена.**] определяется как $X(G)=\sum(\deg_i \cdot \deg_j)^{-1/2}$, где суммирование

ведется по всем ребрам (*i*,*j*) графа *G*. Этот индекс может быть естественным образом обобщен на цепочки длины (*k*-1) следующим образом: ${}^{k}X=\sum$ $(deg_1*deg_2*...*deg_k)^{(-1/2)}$, где суммирование ведется по всем цепочкам, состоящим из *k* атомов, имеющих соответственно степени deg_1, deg_2,...,deg_k.[196]

Индекс Хосойи [248,249] $Z(G)$ задается как $Z(G)=\sum_{k=0}^{m} p(G,k)$, где p(G,k) число способов, с помощью которых *k* ребер графа *G* могут быть выбраны так, чтобы ни какие два ребра не имеют попарно общих концов (по определению p(G,0)=1, p(G,1)=|E(G)|=m). Достоинством предложенного индекса является справедливость реккурентной зависимости, позволяющей разрабатывать эффективные алгоритмы[250] для вычисления этого индекса: p(G,k) = p((G\e),k) + p((G°e),k); $Z(G) = Z((G\backslash e)) + Z((G°e))$; где e - произвольное ребро графа, (G\e) - граф G с удаленным ребром e, (G°e) - граф G с удаленным ребром e вместе со всеми вершинами, инцидентными e. В случае, если *G* является деревом, т.е. связным ациклическим графом, $p(G,k)$ с точностью до знака совпадает с коэффициентом характеристического многочлена матрицы смежности *A*[251]. Индекс Хосойи хорошо коррелирует с температурой кипения алканов, т.к. отражает тенденцию к увеличению этой температуры в зависимости от размера молекулы и степени ее разветвленности [251].

Индекс Ловиша-Пеликана представляет собой наибольшее собственной значение матрицы смежности $A(G)$ и был введен как мера количества разветвлений М-графа [187].

Теоретико-информационные индексы. Разработка теоретико-информационных инвариантов графа является важным направлением в описании особенностей М-графов[252]. В основе построения информационных индексов М-графа[253] лежит разбиение его на элементы (фрагменты, подструктуры), которые затем на основе соотношения эквивалентности группируются в непересекающиеся множества, и для расчета информационного содержания М-графа используется формула Шеннона [254].

Так, пусть множество *S* (множество фрагментов), состоящее из *n* элементов, разбивается на *h* непересекающихся подмножеств S_i мощности k_i (i=1,2,...,h, $\sum_{i=1}^{h} k_i = n$) на основе отношения эквивалентности, определенного на *S*. Тогда этому распределению может быть сопоставлена вероятностная схема:

$S_1, S_2,...,S_h$ $\qquad\qquad$ $\sum_{i=1}^{h} p_i = 1$; (p_i>0); i=1,...,h;

$p_1, p_2,..., p_h$: $\qquad\qquad$ $p_i = (k_i/k)$;

где p_i вероятность того, что выбранный случайным образом элемент множества ***S*** будет находиться в *i*-ом подмножестве $\boldsymbol{S_i}$. Энтропия (или сложность) молекулярной структуры ***G***, связанная с данным видом разбиения, расчитывается по формуле Шеннона [254]: $\boldsymbol{IC(G)} = -\sum_i (p_i * log(p_i))$;

Таким образом, мера сложности ***IC*(*G*)** структуры М-графа ***G*** существенно зависит как от способа разбиения графа (т.е. от способа формирования из ***G*** элементов множества ***S***), а также от *введения на* ***S*** *соотношения эквивалентности*, которое и определяет формирование подмножеств $\boldsymbol{S_i}$ из элементов множества ***S***. Информационное содержание М-графа может рассматриваться как количественная мера его структурной неоднородности или разнообразия. Так, из двух графов $\boldsymbol{G_1}$ и $\boldsymbol{G_2}$ с одинаковым числом вершин граф $\boldsymbol{G_1}$ является более сложным, чем $\boldsymbol{G_2}$, если множество вершин графа $\boldsymbol{G_1}$ разбивается на большее число ***h*** классов эквивалентности, чем для графа $\boldsymbol{G_2}$. Это может быть, например, в том случае, если граф $\boldsymbol{G_2}$ является "более симметричным", чем граф $\boldsymbol{G_1}$. Теоретико-информационные индексы можно рассматривать как количественную меру симметрии соответствующих структур. Зная величину информационного содержания ***IC*** графа ***G***, можно расчитать полное информационное содержание ***TIC***, структурное информационное содержание ***SIC***, информационное содержание связывания ***BIC*** и комплементарное информационное содержание [187]:

$$\boldsymbol{TIC=n*IC;\ SIC=IC*log(n);\ BIC=IC*log(m);\ CIC=log(n)-IC.}$$

1.4.3 Использование мер подобия молекул.

Современное состояние исследований молекулярного подобия связано с использованием понятия близости молекул [255] как для поиска ККСС/QSAR моделей, так и для проведения поиска структурно близких молекул в больших фактографических СБД. Это направление характеризуется резко возросшим в последние годы потоком литературы, посвященной этой весьма актуальной теме[256,257,258,259,260,261,262]. Известно, что математические модели "структура-свойство" строятся, как правило, в рядах структурно родственных соединений, составляющих обучающую СБД. Важной областью применения таких моделей является их использование для внеэкспериментального компьютерного скрининга существующих фактографических СБД с целью поиска молекул, обладающих заданным свойством, экспериментальное значение которого не измерялось и информация о котором отсутствует в СБД. Такая ситуация характерна для СБД, содержащих сведения о уже синтезированных веществах, биологические испытания и экспериментальное измерение активности которых сопряжено со значительными затратами. Кроме этого, модели "структура-свойство" являются основой селекции новых структур при молекулярном дизайне, одно из направлений которого состоит в компьютерной генерации новых структур (на топологическом или пространственно-геометрическом уровне) в заданных структурных ограничениях [263]. Основная математическая и методологическая проблема использования моделей "структура-свойство" при скрининге больших и структурно разнородных СБД заключается в том, что *при построении таких моделей в них не включаются ограничения на область допустимы структур для данной модели* - предполагается, что за этим должен следить химик-исследователь. И действительно, если химик работает в некотором химическом ряду соединений, информация о котором содержится в его локальной СБД, то у него не возникает проблем с использованием моделей "структура-свойство" при конструировании молекул в заданном химическом классе. В последнее время ситуация существенно изменилась. В связи с распространением лазерных компакт дисков и телекоммуникационных средств для обмена химической структурной информацией значительно расширились возможности для поиска веществ, обладающих заданным свойством и относящихся к различным химическим классам. Это относится,

например, к фармакологической активности и токсичным веществам. Возникает задача корректного использования существующих QSAR-моделей при скрининге СБД, содержащей молекулы различных химических классов. Аналогичная ситуация возникает и при оценке физико-химических свойств веществ. Так, известно, что количественные корреляции "структура-свойство", полученные с использованием топологических индексов для свойств алканов, теряют прогностическое качество при значительном возрастании числа атомом углерода в молекуле. Ведение мер (структурного, электростатического, стерического) подобия позволяет провести кластеризацию СБД, описать кластеры и в каждом из них иметь отдельную QSAR модель.

Системы поиска подобных молекул, структуры которых имеют заданную пространственную ориентацию, разрабатываются в ряде крупных фирм, занимающихся конструированием новых биологически активных молекул. Необходимо отметить систему поиска пространственно подобных молекул, созданную в Национальном Институте Рака (США) на основе СБД, накапливаемой с середины 1950-х годов и содержащей в настоящее время более 400 тысяч структур [73]. Инструментальные системы для построения пространственных мер подобия разрабатываются в американских фирмах Tripos и Molecular Simulation Inc (MCI).

1.5 Выводы

Описаны существующие подходы к оценке физико-химических и биологических свойств молекул. Квантово-химические и геометрические методы были рассмотрены как альтернатива методам распознавания образов, получившему название QSAR моделирование. При характеристике того или иного подхода описаны два ключевых момента QSAR модели:

1.Какие признаки использованы для описания молекулы?

2.В каком классе функций построена QSAR зависимость?

Проведено отдельное рассмотрение QSAR моделей физико-химических и биологических свойств, поскольку для последних существенным моментом является необходимость трехмерного представления структур. Описаны различные подходы формирования дескрипторов молекул СБД в виде вектора признаков. Обзор позволяет сделать вывод: использование фрагментов

молекул (подструктур, функциональных групп и т.п.) как признаков при решении ККСС-задач находит широкое применение как в аддитивных схемах оценки физико-химических свойств, так и для конструирования различных топологических индексов. При предсказании биологической активности фрагменты молекул обобщаются до уровня "точек функционального взаимодействия" или областей, обладающих специфическим профилем заданных локальных физико-химических свойств. С другой стороны, для того, чтобы структура обладала биологической активностью между фрагментами в молекуле должны существовать заданные *отношения взаимного пространственного расположения* (3D фармакофор).

Таким образом, является оправданным и перспективным разработка унифицированной методологии QSAR моделирования, основанной на *использовании структурных дескрипторов* различных типов. Для определения теоретических основ построения и кодирования признаков, отражающих число повторения заданного фрагмента в молекуле, весьма перспективным является проведение аналогии между изображением (как структурным объектом распознавания образов) и химической структурой, представленной молекулярным графом. Результаты сопоставления методов описания изображений и молекул приводят к понятию структурного спектра заданной сложности, что обсуждается во второй главе.

2. Глава 2. Формирование структурных дескрипторов молекулярных графов.

В главе рассмотрена постановка QSAR-задачи (т.е. задачи поиска количественных корреляций "структура-свойство") как классической задачи распознавания образов со стандартной информацией. Известно, что задача распознавания образов со стандартной информацией может быть оптимально решена в рамках алгебраического подхода с помощью существующих пакетов программ анализа данных и построения алгоритмов классификации [264].

Характерной особенностью исходной постановки задачи распознавания, решаемой в работе, является представление анализируемых объектов в виде простых помеченных графов, а не в виде заранее фиксированного вектора признаков. *Незаданность признакового пространства*, в котором описываются объекты, подлежащие распознаванию, приводит к постановке задачи формирования такого пространства признаков, в котором может быть получено наилучшее решение. Подобная ситуация с незаданностью векторного представления объектов существует при *анализе и распознавании изображений*. Последняя задача может быть решена в рамках дескриптивного подхода к распознаванию изображения [265]. Проведение аналогии между распознаванием изображений и распознаванием молекул (молекулярных графов) приводит к определению индуктивных структурных спектров графов, используемых для формирование и селекции семейств признаковых пространств. Описана общая схема выбора признакового пространства, адекватного для анализа заданного молекулярного свойства в рамках линейной QSAR-модели.

Признаками графа при его векторизации являются инварианты графа различных типов[266]. Рассмотрена модификация известной общей схемы формирования инвариантов. В ее основе лежит введение алфавита примитивов описания графов. Мощность алфавита, формируемого по данной обучающей структурной БД (СБД), существенно зависит от определения отношения эквивалентности на элементах примитивов. Рассмотрены правила формирования различных отношений эквивалентности путем дополнительного символьного маркирования верших графов. Инвариантами *первого*

уровня является число повторений *примитивов* в графах обучающей СБД; структурными инвариантами *второго уровня* - число повторений *пар примитивов*, находящихся в графе на данном расстоянии; структурными инвариантами *третьего (четвертого) уровня* - число повторений *троек (четверок) примитивов*, расположенных в графе на заданных расстояниях или находящихся относительно друг друга в заданном отношении. Например, могут быть введены "правые" и "левые" структурные инварианты четвертого уровня для упорядоченной пространственной четверки точек, необходимые при анализе хиральных молекул, т.е. структур, переходящих друг в друга при зеркальном отражении, но не накладывающихся при перемещении. Отбор инвариантов, адекватных для прогнозорования заданного свойства, проводится на основе эволюционного подхода - из семейств формируемых шаг за шагом QSAR уравнения выбираются лучшие, которые по построению содержат в себе искомый набор признаков.

Принципиально важной особенностью предложенной методологии является то обстоятельство, что при проведении векторизации помеченных графов и формировании матрицы "граф-признак" X столбцы этой матрицы *получают символьные имена*, которые представляют собой канонические коды подструктур примитивов. Наличие этих имен позволяет осуществлять оптимизацию символьного представления структурных дескрипторов различных уровней на основе однажды сформированной матрицы **X.** Доказана *теорема о расщеплении* столбцов матрицы “граф-признак”, формируемой по обучающей СБД, при проведении дополнительной маркирования (классификации) примитивов описания. Следствием применения теоремы являются ряд схем по формированию оптимальной (в заданных условиях) классификации примитивов и оптимального разбиения на интервалы расстояний при перечислении структурных инвариантов второго, третьего и четвертого уровней.

Описаны алгоритмы анализа молекулярных графов, позволяющие формировать структурные инварианты различных уровней, а также решать задачу поиска графа, инвариантного данному в списке графов на основе символьного построения канонического кода графа.

2.1 ***Задача построения количественных корреляций "структура-свойство" как задача распознавания образов.***

Постановка задачи. QSAR-задачу прогнозирования свойств химических соединений можно рассматривать как *особый случай задачи распознавания образов с учителем со стандартной информацией*. [276] Пусть задана обучающая выборка (структурная БД) ***TS***, которая состоит из **N** объектов и представляет собой список пар $TS=\{(X_1,C_1),(X_2,C_2),...,(X_N,C_N)\}$, где **i**-й объект представлен в виде вектора-строки $X_i=(X_{i1},X_{i2},...,X_{iM})$; C_i - внешний признак, задающий принадлежность **i**-го объекта к одному из классов. Требуется построить классифицирующую функцию ***F*** такую, что на объектах обучающей выборки ее значения приближают внешнюю классификацию **C**, заданную учителем:

$$F(X_i) = C_i + e_i; \quad \sum e_i^2 \rightarrow \min ; \quad i=1,\ldots, N; \qquad (2.1)$$

Уравнение (2.1) используется для классификации новых объектов, не принадлежащих обучающей выборке. Особенности химической интерпретации этой общей постановки задачи распознавания образов информацией состоят в следующем:

1. Внешний признак **C**, задающий классы обучающей СБД, может быть действительным **числом**, определяющим, например, физико-химическое свойство химического соединения;

2. Объекты обучающей СБД представляют собой молекулярные структуры, которые могут рассматриваться как **простые графы**, **имеющие метки** на ребрах и вершинах (молекулярные графы);

3. Классифицирующая функция $F(X_1,X_2,...,X_K)$ традиционно имеет линейный вид:

$$F(X_1,X_2,...,X_K) = b_0 + b_1X_1 + b_2X_2 + ... + b_KX_K ; \qquad (2.2)$$

4. Заранее **не задано** векторное представление молекулярных структур;

Определение. *Молекулярный граф* (М-граф) **G=(V(G),E(G))** - это простой помеченный граф [267], вершины которого (множество **V(G)**) интерпретируются как *атомы молекулы*, а ребра (множество **E(G)**) - как *валентные связи*, соединяющие пары атомов.

Рассматриваемые в работе М-графы имеют следующие специфические особенности:

(1) метками вершин являются *имена атомов* согласно периодической таблице Менделеева - **C**, **O**, **N**, **CL**, **BR**, **F**, **S**,...;

(2) метки ребер обозначают *тип химических связей* - одинарная, двойная, тройная, ароматическая;

(3) атом водорода не считается вершиной М-графа - "водород стерт";

(4) степень вершин не превосходит шести;

(5) каждой вершине приписаны три действительных числа **(X,Y,Z)** - *атрибуты* положения атома в некоторой декартовой системе координат. Расстановка координат проводится так, что длины связей приблизительно равны “условной единице” - “длинные” ребра не допустимы.

(6) число вершин в графе не превышает некоторую константу. Данное условие позволяет ограничиться рассмотрением только низкомолекулярных структур, и не рассматривать, например, белковые молекулы. На практике значение ограничивающей константы равно 255.

Два М-графа $\mathbf{G_1}$ и $\mathbf{G_2}$ *изоморфны* ($\mathbf{G_1} \sim \mathbf{G_2}$), если существует взаимно однозначное соответствие их вершин такое, что сохраняется соответствие ребер, а *метки соответствующих вершин совпадают*. Совпадение *атрибутов* вершин является необязательным.

Векторизация графа. Пусть задан способ вычисления инвариантов $\boldsymbol{F=\{f_1, f_2, \dots, f_m\}}$ М-графа **G** :, $\mathbf{x_1=f_1(G), x_2=f_3(G), \dots, x_m=f_m(G)}$.

$$\mathbf{G \dashrightarrow (x_1, x_2, \dots, x_m)},$$

где $\{\boldsymbol{f_j, j=1,\dots,m}\}$ - алгоритмы вычисления инвариантов - численных значений, совпадающих для изоморфных графов. Тогда по обучающей СБД $\mathbf{\{G_i, i=1,\dots,N\}}$ будет однозначно сформирована матрица (**N x m**) “граф-признак” **X.** Таким образом, приходим к следующей постановке задаче. Пусть в компьютер введена СБД, содержащая **N** записей. Каждая **i**-я запись СБД содержит пару $\mathbf{(G_i, u_i)}$, где $\mathbf{G_i}$ - М-граф, представленный таблицей связности атомов (с

соответствующей символьной меткой атомов и связей); $\mathbf{u_i}$ - численной значение экспериментального свойства вещества, молекулы которого представлены М-графом $\mathbf{G_i}$. Необходимо построить правила QSAR-оценки свойства в виде линейной функции **f** от некоторых инвариантов $\mathbf{\{X_1,X_2,X_3,...X_k\}}$, вычисляемых на М-графе $\mathbf{G_i}$, так, чтобы минимизировать сумму квадратов ошибки:

$$u_i = f(G_i) + \varepsilon_i = \{b_0 + \sum_{j=1}^{K} b_j X_j(G_i)\} + \varepsilon_i, \quad \text{i=1,...N} \qquad (2.3)$$

$$\sum_{i=1}^{N} \varepsilon_i^2 \to \min; \qquad (2.4)$$

где **N** - размер СБД; ε_i - ошибка оценки свойства **u** для **i**-го графа, $\mathbf{X_j(G)}$ - значение **j**-го инварианта графа **G**; $\mathbf{b_j}$ - коэффициенты, определяемы из условия (2).

Замечание 1. В практических приложениях запись СБД может содержать значение не одного свойства u, а нескольких экспериментальных значение различных свойств: $\mathbf{(G_i,u_i,v_i,...,w_i)}$. Для некоторых свойств для одного и того же графа **G** в СБД может содержаться несколько различных экспериментальных значений одного свойства: $\mathbf{(G_i; u_i^1,u_i^2,...u_i^{ku};v_i;...; w_i^1,w_i^2, w_i^3,..., w_i^{kw})}$. Проблема выбора одного значения экспериментального свойства для построения QSAR-оценки будет рассмотрена ниже. Не ограничивая общности, рассмотрим схему выделения QSAR-знаний для одного свойства с единственным для каждого М-графа экспериментальным значением. Для каждого дополнительного свойства приводимая схема вычислений может быть выполнена независимо, а база знаний будет содержать идентификацию свойства для каждой хранимой QSAR-модели.

Замечание 2. Выбор линейной QSAR-функции (2.3) в качестве базовой модели для оценки свойства не является в общем случае обязательным. В настоящей работе этот выбор обусловлен рядом теоретических гипотез об аддитивной природе многих физико-химических свойств, к оценке которых применяется описываемая схема.

Замечание 3. Использование МНК-критерия (2.4) является одним из вариантов оптимизации функции (2.3). Так, для построения “прогностичных”

QSAR-функций могут быть использованы “внешний” критерий[268,269] или критерий “скользящего контроля” [270].

Векторизация М-графов сводит QSAR-задачу к задаче распознавания образов со стандартной информацией. В этом случае для поиска функции классификации (2.1) и решения QSAR-задачи могут быть успешно использованы существующие программные системы [271,272,273]. Действительно, современное состояние математических методов распознавания образов и методов анализа данных позволяет утверждать, что если в матрице “объект-признак” **X** имеются зависимости между вектором внешней классификации **C** и столбцами матрицы, то они будут найдены в рамках тех или иных моделей распознавания[274]. Кроме этого, развиваемый Ю.И.Журавлевым с сотрудниками алгебраический подход к построению алгоритмов распознавания [275] позволяет синтезировать оптимальный (в заданных ограничениях) алгоритм классификации на основе нескольких эвристических алгоритмов.

Алгебраический подход к распознаванию образов. В настоящее время *распознавание образов* это информационная технология разработки алгоритмических баз знаний [276], которая располагает развитой и в определенном смысле законченной математической теорией, сформированной на базе **алгебраического подхода** [275]. Целью *алгебраического подхода к распознаванию* является получение алгоритма, обеспечивающего выделение из представленных данных всей полезной информации и получение решения, полностью и точно соответствующего "информативности" этой информации. В процессе применения алгебраического подхода существенно используются:

- принцип прецедентности;
- формализация понятия "обобщенной близости";
- автоматизация настройки на задачу;
- принцип коррекции окончательного решения путем расширения базового множества алгоритмов.

Процесс решения задачи распознавания образов носит *трехуровневый* характер. Вначале на первом уровне строится эвристическая модель алгоритма, которая отражает специфику решаемой прикладной задачи. На следующем уровне работа ведется с моделями семейств алгоритмов, порождаемых стандартным образом на основе принципа, выбираемого

эвристически. На третьем уровне искомый алгоритм синтезируется из алгоритмов, принадлежащих разным моделям. Главное отличие алгебраического подхода от других методов распознавания состоит в том, что в них *отсутствует третий уровень*, и значит, нет возможности получать точное решение в общем случае [276].

В общем случае под **образом** понимается структурированное приближенное (частичное) описание (эскиз) изучаемого объекта, причем *частичная определенность описания является принципиальным свойством образа* [277] . Образ допускает рекурсивное описание: символ является образом, список образов (символов) является образом. Списочная запись позволяет использовать одно и тоже представление для описания образа произвольного типа независимо от его содержания [278]. Образ состоит из двух групп символов, представляющих соответственно *переменные* и *постоянные* характеристики объекта описания. Описания служат для установления соответствия образов, т.е. для доказательства их идентичности, подобия, сходства, осуществляемого сопоставлением. Сопоставление описаний образов [279] представляет основную задачу как распознавания образов, так и искусственного интеллекта в целом.

В классических постановки задачи распознавания образ описывается вектором признаков (вектором числовых значений) [280,281,282]. В структурных (лингистических) моделях распознавания в качестве образа выступает высказывание, порождаемое грамматикой, характеризующей класс[283]. Построить классификатор означает разработать (синтезировать) модель, в которой учитывается изменчивость образа в заданном классе на основе выбора из заданного набора признаков подмножества, адекватно характеризующего отдельные классы объектов. Существует *два основных этапа распознавания*:

1. приведение исходных данных к стандартному виду, удобному для распознавания образов - *проведение синтеза формализованного описания объекта* на основе имеющейся исходной информации;
2. *собственно распознавание образов* - преобразование формализованного описания в стандартизированную матрицу ответов.

После представление исследуемых объектов в виде вектора признаков дальнейшее алгоритмическое построение системы классификации проводится в рамках одной из следующих моделей [284]:

Модели, основанные на **принципах разделения**. Эти модели различаются заданным классом поверхностей, среди которых выбирается поверхность, в некотором смысле, наилучшим образом разделяющая элементы разных классов [285].

Статистические модели. Этот тип моделей алгоритмов распознавания основан на применении аппарата математической статистики [281,282,285]. Статистические модели применяются в основном в тех случаях, когда известны вероятностные характеристики классов, например, соответствующие функции распределения.

Модели, использующие **метрику между объектами** или построенные на основе *метода потенциальных функций* [286]. В основе этой модели лежит заимствованная из физики идея потенциала, определенного для любой точки пространства и зависящего от того, где расположен источник потенциала. В качестве функции принадлежности объекта классу используют потенциальные функции - всюду положительные и монотонно убывающие функции расстояний в признаковом пространстве.

Модели вычисления оценок (голосования). Модели голосования основаны на принципе *частичной прецедентности* [275,277]. Анализируется "близость" между частями описаний ранее классифицируемых объектов и объекта, который нужно распознать. Наличие близости служит частичным прецедентом и оценивается по некоторому заданному правилу (посредством числовой оценки). По набору оценок близости вырабатывается общая оценка распознаваемого объекта для класса, который и является значением функции принадлежности объекта классу [287].

Модели, основанные на аппарате алгебры логики. В этих моделях признаки объектов рассматриваются как логические переменные. Описания классов строится в виде формул алгебры логики [288]

Распознавание изображений. Дескриптивный подход. Существует глубокая аналогия между распознаванием (классификацией) молекулярных структур и распознаванием изображений, которая основана на незаданности

стандартного описания исследуемых объектов. Известно, что задача распознавание со стандартной информацией и задача *распознавания изображений различны в столь значительной мере, что последняя* ***не допускает*** *непосредственного использования методов и средств классической теории распознавания*, и требует разработки специального направления теории распознавания, предназначенного для работы с изображениями [276]. Теория распознавания изображений развивается в рамках *дескриптивного подхода* в виде дескриптивной теории распознавания изображений [289]. При этом в процессе распознавания участвуют **иерархия формализованных описаний** и представлений изображений, в частности:

1. модели, относящиеся к *различным морфологическим и масштабным уровням представления*, а также

2. многоуровневые модели, позволяющие в процессе распознавания выбирать и изменять необходимую *степень подробности описания* объектов распознавания.

Возникновение дескриптивного подхода вызвано следующими особенностями организации процесса распознавания изображений [290]:

- Синтез формализованного описания изображения как объекта распознавания становится *самостоятельной задачей*, которая ставится и решается непосредственно в рамках процесса распознавания, а не вне его как это обычно происходит при классическом распознавании со стандартной информацией.

- Формализованное описание изображений должно сохранять преимущества "картинки" как средства представления исходной информации. При синтезе описания должен последовательно проводится *принцип порождения* и *принцип иерахичности структуры* распознаваемого объекта.

- Существует глубокая взаимосвязь между процедурами порождения формализованных описаний изображений и процедурами собственно распознавания - *принцип двойственности*. Так результат распознавания может быть представлен в виде нового формализованного описания, сопоставляемого анализируемому изображению.

• Процесс распознавания изображений реализуется как *траектория преобразования* формализованных описаний, что приводит к необходимости введения средств управления "движением этой траекторией".

Таким образом, основные положения дескриптивной теории распознавания изображений состоят в следующем:

1. Систематическое применяется *принципа порождения* для формирования примитивов описания на каждом уровне представления.

2. Используется двойственность процедур синтеза описаний изображений и процедур распознавания.

3. Организация распознавания проводится по схеме "эвристика - модель - коррекция на множестве моделей"

4. задачи распознавания изображений сводится к задаче распознавания со стандартной информацией.

При анализе пространственной организации изображений ставится две основные задачи [291]:

•построения характерных объектов, выделяемых на изображении, и

•обнаружение изменений параметров, связанных с такими объектами.

В результате строятся обобщенные (укрупненные) объекты и проводится граница между множествами таких характерных объектов. Для построения непроизвольных элементов (примитивов) изображения используется следующая рекурсивная процедура. Исходные данные представляют собой описание, составленное из непроизвольных элементов, выделенных на изображении. Из него выбираются элементы, обладающие некоторым подобием. Они подвергаются процедурам группировки и кластеризации, в результате на их основе строятся примитивы следующего уровня - прямые, кривые и т.п. Многократное повторение этой процедуры приводит к получению *нескольких уровней описания изображения* на основе примитивов соответствующего уровня. Эта общая схема построения иерархии описаний изображений может быть конкретизирована и применена для формирования векторного представления молекул (или М-графов). Главное назначение многоуровневого описания М-графов обучающей СБД состоит в том, чтобы реализовать ***основной принцип QSAR-оптимизации***, рассматриваемый в настоящей работе: *проводить поиск оптимального вида классифицирующей*

(прогнозирующей) функции (2.1) не за счет усложнения ее вида (использование полиномов второй или третьей степени, применение нейронных сетей и т.п.), а *на основе выбора признакового пространства, адекватного исследуемому свойству* ***C***.

2.2 Структурные дескрипторы и структурные спектры

Приложение основных положений дескриптивной теории распознавания изображений для поиска описаний М-графов приводит к введению и использованию при решении QSAR-задач понятий структурного спектра [292,293] и индуктивного структурного спектра как набора фрагментов (подграфов) перечисляемых в графах обучающей СБД.

Использование формул (2.3) и (2.4) для построения QSAR-зависимостей является общепринятой классической схемой. Основной проблемой является выбор класса инвариантов $\{X_1,X_2,X_3,...X_k\}$ в рамках которого производится регрессионный поиск коэффициентов $\{b_j\}$. При построении QSAR -уравнений и их использовании для анализа вновь вводимых структур нами были определены следующие требования к выбору инвариантов М-графов для уравнения (2.3):

1. Структурная интерпретируемость - пользователь химик должен иметь возможность легко понять, какие структурные особенности молекулы представляет дескриптор, включенный в QSAR-модель.
2. Простота вычисления - наличие эффективных алгоритмов вычисления инварианта для произвольного М-графа.
3. Простота расширяемости - возможность дополнительно учитывать локальные особенности молекул (например, способность принимать/отдавать электрон, локальный заряд атома, способность к взаимодействию со специфическими функциональными группами).

Этим требованиям не удовлетворяют ни топологические индексы [294,295], ни информационно-теоретические индексы [296], ни метрические индексы [9] М-графов. Всем вышеприведенным условиям удовлетворяют структурные дескрипторы.

Будем представлять М-граф **G** вектором признаков следующим образом. Пусть задана библиотека (список) ациклических фрагментов М-графов $H=\{h_j\}$, $j=1,...,L$. Тогда *структурным дескриптором* $X[h_j]$ назовем пару (h_j,X_j), где X_j -

число повторений фрагмента $\mathbf{h_j}$ в графе **G**. Фрагмент $\mathbf{h_j}$ входит в граф **G**, если существует такой подграф $\mathbf{G' \subseteq G}$, что $\mathbf{h_j \sim G'}$. Вектор-строку $\mathbf{(X_1,X_2,...,X_L)}$ назовем *структурным спектром* графа **G** относительно ***H*** и обозначим его **SS(G|*H*)**.

Для многих физико-химических свойств веществ в результате многолетних исследований были созданы и проверены на практике систематизированные библиотеки $\mathbf{\{h_j\}}$ фрагментов, для которых были рассчитаны соответствующие весовые вклады $\mathbf{b_j}$ [297,298,299]. Предлагаемая нами методология позволяет порождать (под управлением химика-эксперта) по СБД библиотеки фрагментов заданной сложности и отбирать среди всех возможных фрагментов значимые наборы признаков [300,301]. При этом, библиотеки фрагментов, значимых для данного свойства, порождаются автоматически.

Символьное кодирование фрагментов. Для различения фрагментов будем использовать канонических символьный код, обладающий мнемоническим свойством - по коду подструктуры легко восстановить ее "вид". Кодирование фрагментов $\mathbf{h_j}$ будем проводить с помощью канонической линейной нотации [302] (W-нотации), которая является модифицированной системой Висвессера (WLN) для символьного представления М-графов[303,304]. Для ациклических графов правила WLN кодирования определяет выбор начального атома и правила обхода дерева графа.

Определение. *W-код фрагмента* $\boldsymbol{h_j}$ - есть символьная строка $\boldsymbol{W[h_j]}$, составленная из меток $\boldsymbol{LA_i}$ атомов фрагмента $\boldsymbol{h_j}$, записываемых в порядке обхода вершин фрагмента. Каноническим назовем лексикографически младший код.

W-коды по построению обладает следующим важным свойством:

$$\boldsymbol{W}[\mathbf{h_i}]=\boldsymbol{W}[\mathbf{h_j}] \Leftrightarrow \mathbf{h_i \sim h_j}, \qquad (2.5)$$

т.е. коды фрагментов совпадают тогда и только тогда, если фрагменты изоморфны. Формула (**2.5**) позволяет представлять список фрагментов ***H*** в *виде списка символьных кодов (строк)-и работать только с этими строками*.

Ñòðóêòóðíûå ôðàãìåíòû ïðåäñòàâëÿþòñÿ â êîìïüþòåðå êàíîíè÷åñêèì êîäîì [302] - ñòðîêîé ñèìâîëîâ. Âèä êàíîíè÷åñêîãî êîäà è âîçìîæíîñòü ñîäåðæàòåëüíîãî ñðàâíåíèÿ ôðàãìåíòîâ ïîðåäåëÿþòñÿ ñïîñîáîì "ðàçìåòêè" àòîìîâ.

Ïî ïîðàäêåíèþ äëÿ ïîñòðîåíèÿ ñòðóêòóðíûõ ñïåêòðîâ Ì-ãðàôîâ íåîáõîäèìî çàäàòü ñïèñîê ***H*** ñòðóêòóðíûõ ôðàãìåíòîâ. Íàçîâåì ***k-ôðàãìåíòîì*** àòîìíî÷åñêèé ôðàãìåíò, ñîñòîÿùèé èç **k** àòîìîâ. Ïóòåì ïåðå÷èñëåíèÿ â Ì-ãðàôå **G** âñåõ åãî **k**-ôðàãìåíòîâ côîðìèðóåì ñïèñîê ***H***, êîòîðûé îáîçíà÷èì êàê $\boldsymbol{H}_{\mathbf{k}}[\mathbf{G}]$.

Îïðåäåëåíèå. *Èíäóêòèâíûì ñòðóêòóðíûì **k**-ñïåêòðîì* **ISS(k,G)** íàçîâåì ñòðóêòóðíûé ñïåêòð ãðàôà G, ïîñòðîåííûé îòíîñèòåëüíî $\boldsymbol{H}_{\mathbf{k}}[\mathbf{G}]$:

$$\mathbf{ISS(k,G) = SS(G|}\ \boldsymbol{H}_{\mathbf{k}}[\mathbf{G}]\ \mathbf{)}.$$

Маркирование атомов. Будем формировать метки атомов с использованием символьных маркеров. Присвоим каждому атому $\mathbf{v_i}$ М-графа новую символьную метку $\boldsymbol{LA_i}$, которую будем составлять из *символьных маркеров* $\boldsymbol{Q_i^1}$, $\boldsymbol{Q_i^2}$ ... Получаем, что каждый структурный дескриптор представляет собой подграф, атомы которого специальным образом размечены. Разметка атомов в графе может включать в себя такие характеристики атомов как местоположение в структуре, число соседей и т.д. Маркер характеризует определенное локальное свойство атома $\mathbf{v_i}$ (заряд, валентность и т.д.) [307]. Способы разметки атомов задаются экспертом в зависимости от вида изучаемых структур и их свойств.

Формируемые экспертом метки атомов влияют на вид W-кода фрагмента, который есть последовательность меток, соответствующих атомам при обходе фрагмента. Такой способ символьного кодирования фрагментов молекул имеет ряд достоинств:

1. Строки легко анализируются и обрабатываются программными средствами. Алгоритмические языки имеют разнообразные встроенные средства манипулирования строками символов.
2. Строка символов позволяет естественным образом определить лексикографический порядок так, как это принято в трансляторах алгоритмических языков.
3. Присвоение атому метки, образованной символьными маркерами, можно считать естественным обобщением различения атомов по именам элементов таблицы Д.И.Менделеева.

Экспертное приписывание атомам символьных меток (маркеров) можно считать естественным обобщением исходных имен вершин М-графа. Достоинством предложенных дескрипторов явилась их универсальность, интерпретируемость, возможность автоматической генерации и возможность контроля за генерацией фрагментов со стороны химика-эксперта. Представление химической структуры в виде набора символьных кодов означает переход на текстовый уровень описания, что существенно облегчает дальнейшую обработку СБД при формировании базы знаний "структура-свойство".

2.2.1 Формирование матрицы "молекула-признак" (QSAR-матрицы).

Общую схему формирования матрицы "молекула-признак" (или QSAR-матрицы) можно представить в виде следующего алгоритма.

Шаг 1. Эксперт задает:

- ◊ **k** - сложность фрагментов, которые будут перечисляться в М-графах;
- ◊ символьные маркеры Q^1, Q^2, ..., Q^p , на основе которых проводится дополнительная классификация вершин М-графов и определяются символьные метки $\{LA_i\}$ атомов.
- ◊ Тип структурных фрагментов;

Шаг 2. Проводится маркирование атомов: по обучающей СБД $\{ G_i \}$ формируется новая СБД $\{ G^Q_i \}$, структуры которой содержат маркированные атомы, т.е. всем вершинам М-графа присваиваются новые метки: $G_i \rightarrow G^Q_i$;

Шаг 3. Проводится полное перечисления в каждом графе всех его **k**-фрагментов и строится соответствующий список **W**-кодов найденных фрагментов $H_k[G^Q_i]$. Для каждого графа новой размеченной СБД $\{G^Q_i\}$, формируются индуктивные структурные **k-**спектры:

$$G^Q_i \rightarrow ISS(k, G^Q_i) = SS(G^Q_i \mid H_k [G^Q_i]) \qquad i=1,...,N;$$

Шаг 4. Формируется ***H(k)*** - список **W**-кодов всех **k**-фрагментов, встречающихся в графах $\{G^Q_i\}$:

$$H(k) = \bigcup H_k [G^Q_i]; \qquad M=| H(k) |;$$

Каждый фрагмент получает свой порядковый номер **j, j=1,...,M.**

Шаг 5. Для каждого графа G^Q_i СБД формируется структурный спектр относительно общего списка ***H(k):***

$$G^Q_i \rightarrow SS(G^Q_i \mid H(k)) = (X_{i1}, X_{i2}, \ldots, X_{iM}) \qquad i=1,\ldots,N;$$

Вектор строка $(X_{i1}, X_{i2}, \ldots, X_{iM})$ будет являться **i**-ой строкой матрицы ***X*** "молекула-признак", описывающей обучающую СБД в ограничениях, заданных

CH3 -	- CH-	CH2=	> CH -	- CH =
p=1 b="s"	p=2 b="s"	p=1 b="d"	p=3 b="s"	p=2 b="d"
CH ≡	> C <	> C =	= C =	- C ≡
p=1 b="t"	p=4 b="s"	p=3 b="d"	p=2 b="w"	p=2 b="t"

Рисунок 2.1. Распределение валентности атомов углерода и значения маркеров атома: p - маркер степени атома; b - маркер химической связи; (b="s" - все связи одинарные; b="d" - есть двойная связь; b="t" - есть тройная связь; b="w" - есть две двойные связи).

экспертом (на шаге 1). Вектор-столбец X_j определяет повторение **j**-ого **k**-фрагмента в графах обучающей СБД.

Построение матрицы ***X*** сводит QSAR-задачу к задаче распознавания образов со стандартной информацией. Практическая проблема, возникающая при решении этой задачи, заключается в необходимости обработки *очень большого числа столбцов* (признаков) в матрице ***X***. Практически решение этой проблемы, основанное на использовании методов эволюционного программирования, обсуждается в третьей главе.

2.2.2 Маркирование атомов

Введем символные маркеры, предназначенные для учета топологических и химических особенности атомов в молекулярных структурах. В классической теории в простейшем случае атомы различают на основе двух характеристик: химической индивидуальности и валентности [305]. В случае атома углерода валентности и связи могут быть распределены так, как показано на рисунке 2.1.

Для различения топологических особенностей атомов в графе будем использовать характеристику степени атома (число ребер у соответствующей

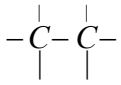
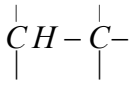
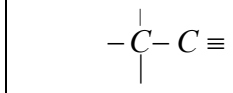
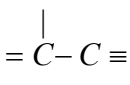
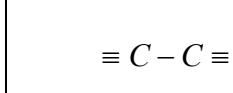
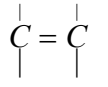
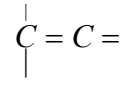
		
“C_4sC_4s”	**“C_3sC_4s”**	**“C_2tC_4s”**
		
“C_3dC_3d”	**“C_2tC_3d”**	**“C_2tC_2t”**
		**=C=C=**
“C_3dC_3d”	**“C_2wC_3d”**	**“C_2wC_2w”**

Рисунок 2.2. Различные связи атома углерода и их кодирование как цепочек в виде "C_pbC_pb", где p, b метки атомов, "C_" - атом углерода.

вершины МХ-графа) и введем соответствующий маркер **p** (p = "power of atom"), который может принимать (для классических органических соединений) семь значений (0, 1, 2, 3, 4, 5, 6). При этом p=0 только для несвязных ("одиночных") атомов в М-графе.

Как видно из рисунка 2.1, одного маркера степени **p** недостаточно для различения всех вариантов "валентного окружения" атома углерода. Введем маркер химической связи атома **b**, который определим следующим образом:

"s" (single) - все связи атома одинарные;

"d" (double) - у атома есть двойная связь;

"t" (triple) - у атома есть тройная связь;

"w" - у атома есть две двойных связи;

"a" (aromatic) - у атома есть ароматическая связь.

Метку атома, в которую включены маркеры, будем записывать в виде строки следующего вида: <имя атома><маркер p><маркер b>.

Проиллюстрируем (рисунок.2.2) разметку атомов и кодирование двухатомных цепочек на примере перечисления возможных связей атома углерода друг с другом (без ароматической связи), построенных с учетом классификации приведенной на рисунке 2.1.

Кроме **p** и **b** маркеров будем использовать еще один маркер **r** (ring) - маркер положения атома в кольцевой системе.

Граф называется связным, если любые его две вершины можно соединить цепью. Будем называть связь (ребро) М-графа кольцевой, если при ее удалении связность графа не нарушается, и цепной (ациклической) - в противном случае. Если атом имеет кольцевые связи, будем называть его "кольцевым атомом". Среди "кольцевых" атомов будем различать "чисто кольцевые" атомы, т.е. атомы, у которых все ребра кольцевые, и "кольцевые с заместителем" атомы, у которых имеется ациклическое ребро.

Определим r-маркер следующим образом:

c (chain) - атом ациклический (цепной);

r (ring) - атом "чисто кольцевой";

s (substitute)- атом "кольцевой с заместителем".

В дальнейшем изложении будем рассматривать эти три маркера как базовые: **p**-маркер степени атома, учитывающий топологическую особенность атома; **b**-маркер, учитывающий особенности химических связей атома и **r**-маркер, учитывающий относительное положение атома в кольцах молекулы. При этом **p**- и **b**-маркеры являются локальными, т.к. они зависят от только от ближайшего окружения атома, а **r**-маркер таковым не является, поскольку зависит от структурной характеристики связи атомов - (кольцевая/цепная связь), которая вычисляется на всей структуре, т.е.

r-маркер является глобальной характеристикой М-графа.

Таким образом, базовая метка атома имеет вид:

<Имя атома><p-маркер><b-маркер><r-маркер> или "**Nnpbr**", где

NN - (2 символа) - имя атома;

p - (1 символ) - p-маркер;

b - (1 символ) - b-маркер;

r - (1 символ) - r-маркер.

Имена атома записываются заглавными символами, p-маркер - цифрой, b- и r- маркеры - строчными буквами. Такая запись позволяет не использовать символ разделителя меток атомов и записывать метки подряд друг за другом (в виде конкатенации символов).

При определении экспертом новых маркеров атомов изменится отношение эквивалентности атомов, а значит и способ описания М-графа в виде набора цепочек. Новые маркеры могут учитывать, стереохимические, электронные или геометрические особенности молекул, быть локальными или глобальными [306]. Способы содержательного определения новых маркеров, введенных экспертом, будут нуждаться в практической проверке путем решения QSAR-задачи (2.3) и (2.4).

ID: 34 **1. (7, C ***)** **2. (8, C ***C ***)** **3. (11, C ***C ***C ***)**	**ID: 57** **1. (8, C ***)** **2. (8, C ***C ***)** **3. (11, C ***C ***C ***)**
ID: 257 **1. (6, C ***)** **2. (5, C ***C ***)** **3. (7, C ***C ***C ***)**	**ID: 339** **1. (5, C ***)** **2. (3, C ***C ***)** **3. (3, C ***C ***C ***)**
7. (2, O *)** **8. (1, O ***C ***O ***)**	

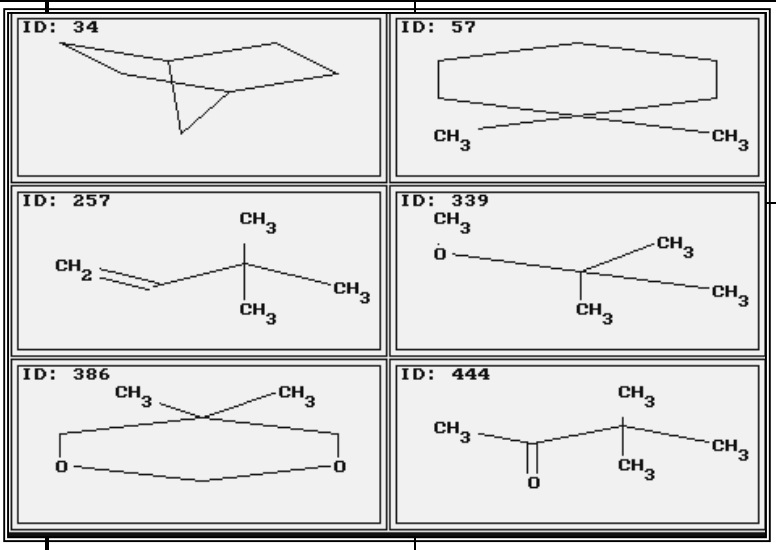


Рисунок 2.3. Структуры модельной СБД, на которой иллюстрируется построения индуктивных структурных спектров.

#1. C ***

#2. C ***C ***

#3. C ***C ***C ***

#4. C ***C ***O ***

#5. C ***O ***

#6. C ***O ***C ***

#7. O ***

#8. O ***C ***O ***

Коды дескрипторов, соответствующие столбцам-признакам матрицы X

Матрица X "граф-дескриптор", метка атомов = "Имя атома", "*" -базовые**

	#1	#2	#3	#4	#5	#6	#7	#8
ID: 34	**7**	**8**	**11**	**0**	**0**	**0**	**0**	**0**
ID: 57	**8**	**8**	**11**	**0**	**0**	**0**	**0**	**0**
ID: 257	**6**	**5**	**7**	**0**	**0**	**0**	**0**	**0**
ID: 339	**5**	**3**	**3**	**3**	**2**	**1**	**1**	**0**
ID: 386	**6**	**4**	**6**	**2**	**4**	**2**	**2**	**1**
ID: 444	**6**	**5**	**7**	**2**	**1**	**0**	**1**	**0**

Пример построения структурных спектров. Рассмотрим модельную СБД, состоящую из шести структур, выбранных из СБД "энтальпия образования" (глава 5). Графическое изображение М-графов приведено на рисунке 2.3, полученном в системе ChemAdd.

Таблица 2.1 Структурный спектр модельной СБД- маркировка "*", k-цепочки (k=1,2,3)**

На следующей странице показано, как меняется структурный спектр при проведении дополнительной классификации атомов - включено маркирование степени атома в М-графе. Число столбцов в матрице **X1** увеличилось с восьми до 35-ти. В матрице **X1** произошло "расщепление" столбцов. Так, первый столбец матрицы **X**, соответствующий дескриптору **"C_***",** и равный **$(7,8,6,5,6,6)^T$** в новой матрице **X1** расщепился на 4 столбца, **"C_1**" (#1); "C_2**" (#14); "C_3**" (#25);"C_4**" (#29)**, которые соответственно равны (см.ниже) **$(0,2,4,4,2,4)^T$, $(5,5,1,0,3,0)^T$, $(2,0,0,0,0,1)^T$, $(0,1,1,1,1,1)^T$.** Их сумма равна первому столбцу исходной матрицы **X**, соответствующей дескрипторам без маркирования атомов.

Далее, будет показано, что такое свойство присуще всем столбцам исходной матрицы (теорема о расщеплении). Это свойство новой матрицы может быть использовано (глава 3) для оптимизации (в автоматическом режиме) кодов дескрипторой. Возможность оптимизации основана на том, что для каждого столбца-признака известен его код. Так, например, для нового дескриптора с обобщенным кодом **"C_Y**"**, где **Y={2,3}**, можно сразу выписать столбец-признак, равный сумме соответствующих столбцов матрицы **X1:**

<"C_Y">(7,5,1,0,3,1)^T=<"C_2**">(5,5,1,0,3,0)^T + <"C_3**">(2,0,0,0,0,1)^T**

где в угловых скобках указаны "имена" (коды) столбцов. Адекватность нового дескриптора **"C_Y**"** исследуемому свойству будет проверена при автоматическом выполнении МГУА расчетов (глава 3).

ID: 34	ID: 57
1, 7, C *s*	1, 8, C *s*
2, 6, C *s*C *s*	2, 7, C *s*C *s*
3, 5, C *s*C *s*C *s*	3, 6, C *s*C *s*C *s*
ID: 237	ID: 339
1, 4, C *d*C 2**C 2**C 3**)	1, 5, (1, C 1**C 4**C 1**)
2, 3, C *d*C 2**C 3**)	2, 4, (2, C 1**C 4**C 2**)
3, 2, C *d*C *s*C 2**)	3, 3, C *s*C *s*C *s*
4, 1, C *d*C *s*	4, 2, C *s*C *s*O *s*
5, 3, C *d*C *s*C *s*	5, 1, (3, C 2**C 2**C 2**)
6, 4, C *s*	6, 1, C *s*O *s*C *s*
7, 3, C *s*C *s*	7, 1, O *s*
8, 3, C *s*C *s*C *s*	10. (1, C 2**C 4**C 2**)
	11. (1, C 4**)
ID: 386	ID: 444
1, 6, C *s*	1, 3, C *d*
2, 4, C *s*C *s*	2, 2, (4, C 1**)
3, 5, C *s*C *s*C *s*	3, 3, C *s*C *s*
4, 3, C *s*C *s*C 2**)	4, 1, (3, C 1**C 4**C 1**)
5, 4, C *s*O *s*	5, 2, C *s*C 1**C 4**O 2**)
6, 2, C *s*O *s*C *s*	6, 1, C *s*C *d*C *s*
7, 2, O *s*	7, 2, C *s*C *d*O *d*
8, 1, O *s*C *s*O *s*	8, 1, (3, C 4**)
	9, 3, C *s*C *s*C *s*
	10, 1, O *d*

ID: 386	ID: 444
1. (2, C 1**)	1. (4, C 1**)
2. (2, C 1**C 4**)	2. (1, C 1**C 3**)
3. (1, C 1**C 4**C 1**)	3. (1, C 1**C 3**C 4**)
4. (4, C 1**C 4**C 2**)	4. (1, C 1**C 3**O 1**)
5. (3, C 2**)	5. (3, C 1**C 4**)
6. (2, C 2**C 4**)	6. (3, C 1**C 4**C 1**)
7. (1, C 2**C 4**C 2**)	7. (3, C 1**C 4**C 3**)
8. (4, C 2**O 2**)	8. (1, C 3**)
9. (2, C 2**O 2**C 2**)	9. (1, C 3**C 4**)
10. (1, C 4**)	10. (1, C 3**O 1**)
11. (2, C 4**C 2**O 2**)	11. (1, C 4**)
12. (2, O 2**)	12. (1, C 4**C 3**O 1**)
13. (1, O 2**C 2**O 2**)	13. (1, O 1**)

столбец	дескриптор
#1	C *d*
#2	C 1**
#3	C *d*C *d*
#4	C 1**C 2**
#5	C *d*C *d*C *s*
#6	C 1**C 3**
#7	C *d*C *s*
#8	C 1**O 1**
#9	C 1**C 4**
#10	C *d*C *d*C *s*
#11	C 1**C 4**C 2**
#12	C *s*C 1**O 2**
#13	C 1**O 2**C 4**
#14	C *s*C 2**C *d*
#15	C 2**C 2**
#16	C *s*C 2**C 2**C 2**
#17	C 2**C 2**C 3**
#18	C *s*C 2**C 2**C 4**
#19	C 2**C 3**
#20	C *s*C 2**C 3**C 2**
#21	C 2**C 4**
#22	C *s*C 2**C 4**C 2**
#23	C 2**O 2**
#24	C *s*C 2**O 2**C 2**
#25	C 3**
#26	C *s*C *s*C *s*
#27	C 3**C 4**
#28	C 3**O 1**
#29	C 4**
#30	C *s*C *s*
#31	C 4**C 2**O 2**
#32	C *s*C *s*O *s*
#33	C 4**O 2**
#34	O 1**
#35	O 2**
	O 2**C 2**O 2**

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20
0	0	0	0	0	0	0	0	0	0	0	0	0	5	2					
2	0	0	0	0	0	2	1	4	0	0	0	0	5	4					
4	1	1	0	0	0	3	3	3	0	0	0	0	1	0					
4	0	0	0	0	0	3	3	0	0	3	1	1	0	0	0	0	0	0	0
2	0	0	0	0	0	2	1	4	0	0	0	0	3	0	0	0	0	0	0
4	0	0	1	1	1	3	3	0	3	0	0	0	0	0	0	0	0	0	0

#21	#22	#23	#24	#25	#26	#27	#28	#29	#30	#31	#32	#33	#34	#35
0	0	0	0	2	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0	1	0
2	1	4	2	0	0	0	0	1	2	0	0	0	2	1
0	0	0	0	1	0	1	1	1	0	1	0	1	0	0

Коды дескрипторов, соответствующие столбцам-признакам

Матрица X 1
"граф-дескриптор",
метка атомов = "NNp"**

Структурный спектр - маркировка "*b*", k-цепочки (k=1,2,3)

Коды дескрипторов, соответствующие столбцам-признакам матрицы X2

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17
0	0	0	0	0	0	7	0	0	8	11	0	0	0	0	0	0
0	0	0	0	0	0	8	0	0	8	11	0	0	0	0	0	0
2	1	1	1	3	0	4	0	0	3	3	0	0	0	0	0	0
0	0	0	0	0	0	5	0	0	3	3	3	2	1	0	1	0
0	0	0	0	0	0	6	0	0	4	6	2	4	2	0	2	1
1	0	0	2	3	1	5	1	2	3	3	0	0	0	1	0	0

Матрица X2 " граф-дескриптор", метка ='NN*b*"

В матрице **X2** число столбцов увеличилось с восьми до 17-ти. Снова произошло "расщепление" столбцов. Так, первый столбец, соответствующий дескриптору **"C_***", и равный **$(7,8,6,5,6,6)^T$** в новой матрице расщепился на 2 столбца, **"C_*d*" (#1); "C_*s*" (#7)**, которые соответственно равны **$(0,0,2,0,0,1)^T$, $(7,8,4,5,6,5)^T$.** Их сумма равна первому столбцу исходной матрицы **X**,соответствующей дескрипторам без маркирования атомов. Расщепление происходит автоматически согласно введенной маркировки атомов и алгоритмы построения матрицы "граф-дескриптор".

Для рассматриваемой модельной СБД при маркировке атомов "NN*b*" в матрице **X2** были получены "вырожденные" признаки, т.е. признаки, уникальные для данной молекулы. Им соответствуют столбцы **(#2); (#3); (#6); (#8); (#9); (#15); (#17)** матрицы **X2**. Легко видеть, что объединение столбцов **<#3; C *d*C *d*C *s*>** и **<#8; C *s*C *d*C *s*>** дает новый невырожденный дескриптор с кодом **C ***C *d*C *s = $(0,0,1,0,0,1)^T$**. Т.е. произошло обобщение маркера первого атома цепочки.

Общая методика экспертного конструирования новых маркеров состоит в применении следующего алгоритма [300,301]. Пусть на атомах М-графа определена некоторая числовая функция **φ(v),** которая может представлять собой или локальный топологический инвариант вершины[307,308], или физико-химическое свойство атома (с учетом его окружения). Пусть новый маркер **m** должен (по мнению химика-эксперта) принимать, например, три символьных значения: **m = (a,b,c).** Разделим область определения **φ(v)** на три непересекающихся интервала: **m1, m2, m3**, которые покрывают всю область значений функции **φ(v),** и определим **m**-маркер:

"a", если **φ(v)** принадлежит интервалу **m1**;

"b", если **φ(v)** принадлежит интервалу **m2;**

"c", если **φ(v)** принадлежит интервалу **m3;**

где **φ(v) -** значение маркерной функции на **v**-атоме М-графа. Разбиение области определения на интервалы **m1, m2, m3** должно задаваться:

- Содержательной интерпретацией маркерной функции **φ(v),**
- Характером исследуемого свойства.

Если новый **m**-маркер включен в формирование символьных меток атомов, то при автоматическом перечислении и **W**-кодировании неизоморфных фрагментов будет сформирован новый список первичных **k-**фрагментов. Адекватность этого набора **k-**фрагментов (для оценки исследуемого свойства молекул) будет проверена на основе построения и оценки качества QSAR-моделей. Для формирования маркера могут использоваться формальные топологические свойства вершин графа, заряды, дистанции, локальные стерические свойства, на основании которых строятся многие известные топологические и информационные индексы.

В качестве "маркерной функции" возможно использование квантово-химических характеристик [309,310]: заряд на атоме и поляризуемость атома, индекс свободной валентности, порядки связей по Уайбергу и Коулсону, спиновые плотности. электронно-топологические индексы[311,312]. Например, для маркировки атомов могут быть использованы “индекс внутреннего состояния” ***I*** и “индекс электротопологического состояния атома” S_i: [307,313]:

$$I = (d^n+1)/\,d,\ d^n = Z^n - h,$$

$$S_i = I_i + D\,I_i, \quad D\,I_i = \sum_j (I_i - I_j)/r_{ij}^2$$

где: d^n - количество валентных электронов атома в скелете молекулы, Z^n - число валентных электронов атома, h - число связанных с ним атомов водорода, d - степень атома, r_{ij} - число атомов в кратчайшем пути от атома i до атома j.

Таким образом, метод конструирования маркера Q основан на задании функции *локального свойства* $\varphi(v)$, вычисляемой на атомах М-графа. Применение *операции* ξ *разбиения* области значений функции $\varphi(v)$ на непересекающиеся интервалы позволяет приписать каждому интервалу символьное значение маркера Q. *Пара (φ,ξ)задает экспертную классификацию атомов*. Выбор локального свойства φ является определяющим моментом для последующего индуктивного построения признаковых пространств, среди которых будет проводится селекция пространства, адекватного для анализируемого свойства молекул.

2.2.3 Свойства структурных спектров маркированных графов

Лемма 2.1 Пусть граф $\mathbf{G^Q}$ получен из графа **G** **Q**-маркированием его атомов $\mathbf{Q}=(\varphi,\xi)$: $\mathbf{G} \to \mathbf{G^Q}$ и пусть фрагмент $\mathbf{h_p^Q} \in \mathbf{ISS(k, G^Q)}$. Рассмотрим *операцию* Δ_Q *стирания* маркера **Q** в коде $\mathbf{W}[\mathbf{h_p^Q}]$ фрагмента $\mathbf{h_p^Q}$. Тогда существует фрагмент $\mathbf{h_j} \in \mathbf{ISS(k,G)}$, такой что $\Delta_Q(\mathbf{W}[\mathbf{h_p^Q}])=\mathbf{W}[\mathbf{h_j}]$, т.е. код $\mathbf{W}[\mathbf{h_j}]$ фрагмента $\mathbf{h_j}$ совпадает с кодом $\mathbf{W}[\mathbf{h_p^Q}]$ после стирания **Q**-маркера.

Доказательство. Установим взаимно однозначное соответствие между вершинами и ребрами графа $\mathbf{G^Q}$ и графа **G.** Поскольку фрагмент $\mathbf{h_p^Q} \in \mathbf{ISS(k, G^Q)}$, то $\mathbf{h_p^Q}$ принадлежит графу $\mathbf{G^Q}$, т.е. существует подграф $\mathbf{H^Q}$ в графе $\mathbf{G^Q}$, изоморфный $\mathbf{h_p^Q}$. Рассмотрим соответствующий подграф **H** в немаркированном графе **G.** По определению маркирования метки на вершинах подграфов $\mathbf{H^Q}$ и **H** будут различаться только символом маркера, поэтому стирание маркера в W-коде $\mathbf{W}[\mathbf{h_p^Q}]$ эквивалентно стиранию маркера на всех вершина подграфа $\mathbf{H^Q}$, что делает подграфы $\mathbf{H^Q}$ и **H** изоморфными, т.е. их W-коды будут совпадать согласно (2.3). Таким, образом подграф **H** и является искомым фрагментом.

Лемма 2.2 Пусть { $\mathbf{h_p^Q}$ }, ***p=1,...,P***; - множество фрагментов из **ISS(k,G^Q)** таких, что $\Delta_Q(\mathbf{W}[\mathbf{h_p^Q}])=\mathbf{W}[\mathbf{h_j}]$, где $\mathbf{h_j} \in$ **ISS(k,G)**, тогда $\mathbf{X}[\mathbf{h_j}]= \sum_{p=1,P} \mathbf{X}[\mathbf{h_p^Q}]$ **,** где $\mathbf{X}[\mathbf{h_j}]$ - число повторений фрагмента $\mathbf{h_j}$ в графе **G**.

Доказательство. Для фрагмента немаркированного графа $\mathbf{h_j} \in$ **ISS(k,G)** условие $\Delta_Q(\mathbf{W}[\mathbf{h_p^Q}])=\mathbf{W}[\mathbf{h_j}]$ означает, что каждый фрагмент $\mathbf{h_p^Q}$ маркированного графа получен из фрагмента $\mathbf{h_j}$ с помощью дополнительной дифференциации вершин на основе введение маркера. Поэтому число повторение немаркированного фрагмента $\mathbf{h_j}$ будет совпадать с общим числом повторений маркированных фрагментов { $\mathbf{h_p^Q}$ }, которые порождены $\mathbf{h_j}$, что эквивалентно условию $\mathbf{X}[\mathbf{h_j}]= \sum_{p=1,P} \mathbf{X}[\mathbf{h_p^Q}]$, что и требовалось доказать.

Лемма 2.3 Для каждого фрагмента $\mathbf{h_j} \in$ **ISS(k,G)** существует множество фрагментов $\{\mathbf{h_p^Q}\} \subseteq$ **ISS(k,G^Q)**, такое что

$$\Delta_Q(\mathbf{W}[\mathbf{h_p^Q}])=\mathbf{W}[\mathbf{h_j}]; \quad \textbf{\textit{p=1,...,P}} \quad \text{и} \quad \mathbf{X}[\mathbf{h_j}]= \Sigma_{p=1,P} \mathbf{X}[\mathbf{h_p^Q}], \qquad (2.6)$$

Доказательство. Для каждого фрагмента немаркированного графа $\mathbf{h_j} \in$ **ISS(k,G)** дополнительная дифференциация вершин на основе маркера означает получение множества маркированных фрагментов { $\mathbf{h_p^Q}$ } таких, что $\mathbf{X}[\mathbf{h_j}]= \sum_{p=1,P} \mathbf{X}[\mathbf{h_p^Q}]$. Стирание маркера в **W**-коде каждого из этих маркированных фрагментов приводит к получению **W**-кода породившего их фрагмента немаркированного графа $\mathbf{h_j}$, т.е. $\Delta_Q(\mathbf{W}[\mathbf{h_p^Q}])=\mathbf{W}[\mathbf{h_j}]$. Таким образом, каждый фрагмент индуктивного структурного спектра немаркированного графа $\mathbf{h_j}$ после проведения маркирования вершин порождает множество фрагментов в индуктивном структурном спектре маркированного графа, которое удовлетворяет условию (2.4.), что и требовалось доказать.

Таким образом, проведение **Q**-маркирования атомов в графе **G** приводит к "расщеплению" его индуктивного структурного спектра. Будем называть множество фрагментов $\{\mathbf{h_p^Q}\}$, удовлетворяющее условиям (**2.4**), *множеством расщепления* фрагмента $\mathbf{h_j}$ по маркеру $\mathbf{Q}=(\varphi,\xi)$ и обозначим это множество как $\mu[\mathbf{h_j}, (\varphi,\xi)]$. Тогда из лемм 2.1-2.3 непосредственно вытекает:

Теорема 2.1 (о расщеплении). Пусть по обучающей СБД $\{\mathbf{G_i}\}$ построены две матрицы "молекула-фрагмент": матрица $\boldsymbol{X}$ на основе $\mathbf{ISS(k,G_i)}$, и матрица $\boldsymbol{X}^{\mathbf{Q}}$ на основе $\mathbf{ISS(k,G_i^Q)}$, $\boldsymbol{i=1,\ldots,N.}$ Тогда каждый вектор-столбец $\mathbf{X_j=X[h_j]}$ матрицы $\boldsymbol{X}$ равен сумме вектор-столбцов $\mathbf{X_p^Q=X[h_p^Q]}$ матрицы $\boldsymbol{X}^{\mathbf{Q}}$:

$$\mathbf{X[h_j]}= \Sigma_{\mathbf{p=1,Pj}}\ \mathbf{X[h_p^Q]}, \text{ где } \mathbf{h_p^Q} \in \mu[\ \mathbf{h_j}, (\varphi,\xi)];\ \ \boldsymbol{p=1,\ldots,P_j;}\ \ \mathbf{P_j} = |\ \mu[\ \mathbf{h_j}, (\varphi,\xi)]\ |$$

Утверждение 2.1. позволяет построить МГУА-схему автоматического *поиска обобщений Q-маркера* и формирования фрагментов, адаптированных для исследуемого свойства $\boldsymbol{C}$. Методы таких построений рассмотрены в главе 3.

2.3 Несвязные фрагменты и их представление

2.3.1 Примитивы описания молекулярных графов

По аналогии с представлением в ЭВМ изображений, подлежащих распознаванию, применим единообразный подход для формирования структурных дескрипторов, описывающих молекулы. Построение структурных дескрипторов будем проводить индуктивно с постепенным усложнением фрагментов.

Пусть в молекуле перечислены "*примитивы описания*" - "*базовые фрагменты*" или "*особые точки*". На первом шаге индуктивного порождения будем формировать дескрипторы, описывающие *взаимное расположение* пар примитивов в молекуле.

Алгоритм перечисления и кодирования пар примитивом таков:

(1) Молекула последовательно рассматривается на различных уровнях представления [314] как топологический, планарный или пространственный объект;

(2) На каждом уровне представления в молекуле определяются *примитивы (элементы описания)* в виде экспертных алгоритмических правил. Выбор примитивов проводится одним из следующим способов:

- Маркированные атомы (**1**-фрагменты).
- Цепочки маркированных атомов (**k-**цепочки, содержащие **k** атомов **k**>1).

- Точки, не связанные непосредственно с вершинами М-графа (атомами). Такие точки выбираются в пространстве, окружающем молекулу. Они могут быть расположены на молекулярных поверхностях различного вида, что подробно описано в 6-ой главе

(3) Каждый примитив задается своими координатами (планарными или пространственными) и *своим **W**-кодом*. **W**-код примитива составляется из *символьных меток*, формируемых по правилам, заданных экспертом; два примитива считаются эквивалентными (при подсчете числа повторений в молекуле) тогда и только тогда, когда совпадают их **W**-коды.

(4) Для всех примитивов, перечисленных в М-графе, строится *матрица отношений* (например, матрица расстояний) $\boldsymbol{D}=\{ d_{ij} \}$, где d_{ij} - задает отношение между двумя примитивами - например, (топологическое или евклидово) расстояние между **i**-ым и **j**-ым элементами описания.

(5) Определяется способ дискретизации отношения ***D*** - вводится разбиение на интервалы и строится матрица $\boldsymbol{P}=\{ p_{ij} \}$, где p_{ij} - номер интервала, в который попало значение отношения (расстояние) d_{ij};

(6) Путем перечисления всех пар примитивов и их отношений составляется список **W**-кодов несвязных фрагментов h_{ij} в виде:

$$W[\, h_{ij} \,] = "(W_i, W_j, P)"; \qquad (2.7)$$

где W_i и W_j - **W-**коды примитивов, входящих в пару; **P** - символьный код (номер) интервала; $X = \#["(W_i,W_j,P)"]$ - число повторений **W-**кода $"(W_i,W_j,P)"$ в молекуле.

2.3.2 Кодирование несвязных фрагментов

Символьная запись $"(W_i , W_j , P)"$ есть код фрагмента. Будем *по определению* полагать, что *два фрагмента равны, если совпадают их **W**-коды*. Такое задание эквивалентности несвязных фрагментов, состоящих из двух примитивов, позволяет сформулировать следующее обобщение индуктивного структурного спектра.

Структурные дескрипторы второго уровня.Пусть примитивы заданы **k**-фрагментами и пусть построен список записей, следующего вида:

$$(W[2],\ X) = (\ "(W_i , W_j , P)",\ X\ )$$

где **X** - число повторений в молекуле такой подструктуры, которая состоит из двух **k**-фрагментов и имеет **W-**код: **W[2]="(W_i , W_j , P)"** Назовем такой список *индуктивным структурным спектром второго уровня* и обозначим его символом **ISS_2(k,G).** Пару **("(W_i , W_j , P)", X)** назовем структурным дескриптором второго уровня.

Построим **ISS_2(k,G)** для каждой молекулы обучающей СБД (вершинам которых уже присвоены символьные метки):

$$G_i \rightarrow ISS_2(k, G_i), \qquad i=1,...,N;$$

Объединим все индуктивные структурные спектры второго уровня **ISS_2(k,G_i)** и сформируется общий (для всех молекул СБД) список **W**-кодов всех пар **k**-фрагментов, встречающихся в графах СБД: $H_2(k) = \bigcup ISS_2(k, G_i),;$ Каждый фрагмент получает порядковый номер **j**, **j=1,...,M.** $M=|H_2(k)|;$

Для каждого графа G_i СБД сформируем структурный спектр относительно этого общего списка несвязных фрагментов $H_2(k)$:

$$G_i \rightarrow SS(G_i \mid H_2(k)) = (X_{i1},X_{i2},...,X_{iM}) \qquad i=1,...,N;$$

Вектор строка $(X_{i1},X_{i2},...,X_{iM})$ будет являться **i**-ой строкой матрицы X "молекула-признак", описывающей граф G_i обучающей СБД в ограничениях, заданных экспертом. Вектор-столбец X_j определяет повторение **j**-ого несвязного фрагмента в графах обучающей СБД **(j=1,...,M;** $M=|H_2(k)|$**).**

Описанное построение индуктивного структурного спектра второго уровня **ISS_2(k,G).** и соответствующей ему матрицы X "молекула-признак" (описывающей обучающую СБД в ограничениях, заданных экспертом) позволяет сделать следующее важное заключение: *Теорема 2.1 о расщеплении справедливо для $ISS_2(k,G)$ и для соответсвующей матрицы X "молекула-признак".*

Структурные дескрипторы (q+1) уровня (q>2) можно определить индуктивно следующим образом. Пусть примитивы заданы **k**-фрагментами и пусть построен индуктивный структурный спектр **q**-ого уровня **ISS_q(k,G)**

Сформируем список записей, следующего вида:

(W[q+1], X) = ("(W[q]$_i$, W_j , P[q])", X)

где **X** - число повторений в молекуле такой подструктуры, которая состоит из **k**-фрагмента с кодом $\mathbf{W_j}$ и структурнго дескриптора **q**-ого уровня, который имеет код $\mathbf{W[q]_i}$. Отношение в М-графе между **k**-фрагментом $\mathbf{W_j}$ и структурного дескриптора **q**-ого уровня задается отношением **P[q].** Назовем этот список *индуктивным структурным спектром* **(q+1)**-ого *уровня* и обозначим его символом $\mathbf{ISS_{q+1}(k,G).}$ Пару **(W[q+1], X)** назовем структурным дескриптором **(q+1)**-ого уровня.

2.4 Алгоритмы анализа структур молекулярных графов

Для вычисления локальных маркерных функций на атомах, для перечисления и кодирования **k**-фрагментов, сравнения графов и подграфов была разработана библиотека базовых алгоритмов обработки и анализа структур М-графов.

Для представления алгоритмов в граф-ориентированной нотации будем использовать следующие обозначения:

G = (V, E)	- молекулярный химический граф;
V(G), V	**-** список атомов **G-**графа;
E(G), E	- список ребер (пар атомов) **G**-графа;
n = | V |	- число атомов **G**-графа;
m = | E |	- число ребер **G**-графа;
s, u, v	- атомы **G**-графа;
e = (u,v)	- ребро **G**-графа;
v(i), v_i .	- атом, имеющий номер **i;**
Г(v)	- список атомов, смежных с **v**-атомом;
p(v) = |Г(v)|	- степень **v**-атома;
A(v), AR(v)	- числовые метки, приписанные атомам графа;
B(e)	- числовые метки, приписанные ребрам графа;
F, FN	- списки атомов;
[ ]	- пустой список;
F:=F+[v]	- занесение **v**-атома в **F**-список.
F:=F-[v]	- удаление **v**-атома из **F**-списка

Для записи алгоритмов будем использовать ПАСКАЛЬ-подобную нотацию. Например, цикл по всем **v**-атомам из **F**-списка будет выглядеть как

```
for v-atom: (v from F) do
begin . . . . . . . . . . . . end;
```

2.4.1 “Волновой” алгоритм и матрица расстояний графа.

Для перечисления в М-графе цепочек заданной длины используется *"волновой" алгоритм*. Этот же алгоритм с некоторыми модификациями используется во многих разработанных для системы BIBIGON программных

модулях. По своей схеме волновой алгоритм является ускоренным параллельным вариантом алгоритма Дейкстры поиска пути минимальной длины в графе с положительными весами на ребрах [45]. Алгоритм был предложен в работе [46]. Для М-графов, имеющих "единичные веса" на ребрах, действие алгоритма похоже на работу алгоритма "степного пожара", используемого при обработке изображений [46].

Пусть все вершины М-графа имеют дополнительную целочисленную **A**-метку, равную нулю. Зададим (стартовый) **s**-атом, его **A**-метка равна **1**. Стартовый атом волнового алгоритма можно интерпретировать как "точку поджога" графа в начальный момент времени **T=1**. "Волна" распространяется по атомам, имеющих нулевые **A**-метки: в момент времени **T=2** волна достигает соседей стартового атома и переводит их **A**-метки в состояние **2**, затем **(T=3)** - соседей этих соседей и переводит их **A**-метки в состояние **3** и т.д. Таким образом, в **A**-метках фиксируется момент времени прохождения волны через атом. Алгоритм использует два рабочих списка атомов: **F**-список и **FN**-список. **F**-список содержит атомы "текущей" волны, а в **FN**-списке формируются "новая" волна следующего шага. Алгоритм прекращает работу, когда **FN**-список оказывается пустым. Числовая **A**-метка атома фиксирует момент времени прохождения волны через этот атом. Поэтому величина **(A(v)-1)** будет равна длине цепи, соединяющей стартовый **s**-атом и данный **v**-атом. Легко видеть, что более короткой цепи, соединяющей эти атомы, не существует. Это следует из построения волны: в каждый такт времени происходит параллельное (по атомам) распространение волны, т.е. она кратчайшим путем доходит до **v**-атома. Этот факт позволяет сформулировать следующее важное утверждение.

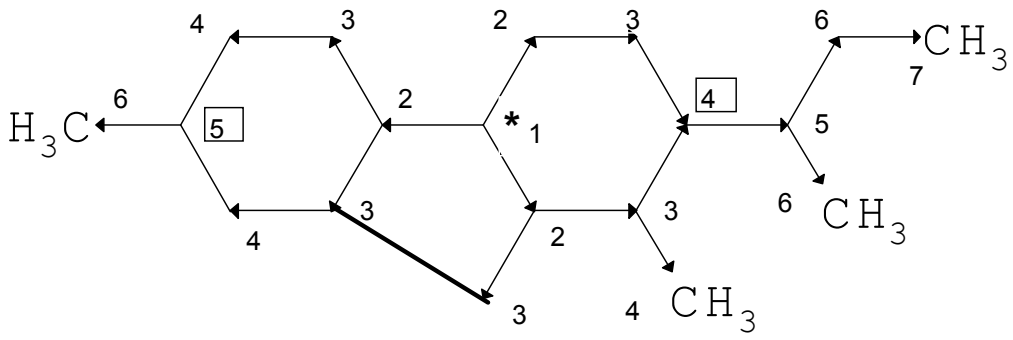


Рисунок 2.5 Волновая разметка графа. Пример "встречи на ребре" и "встречи на атоме". Числами показаны А-метки атомов. А-метки атомов, на которых произошла "встреча" ("встреча на атоме") выделены квадратом. Ребро, на котором произошла встреча ("встреча на ребре"), также изображено жирным.

Утверждение 2.1 Пусть **s** - стартовый атом и проведена волновая разметка графа. Тогда **A**-метки, расставляемые волновым алгоритмом (**A(j)** - **A**-метка **j**-атома), обладают следующим свойством. Значение (**A(j)-1**) равно элементу **D(s,j)** матрицы расстояний графа.

Алгоритм Волна (G, s, A);

Вход: **G** - М-граф: **G = (V, E); s** - стартовый атом, **s**$\in$**V(G**).

Выход: **A** - список числовых меток атомов такой, что $a_i=(d_{si}+1)$, где d_{si} - элемент матрицы расстояний.

Локальные данные:**F, FN** - списки атомов.

```
{Инициализация}
for v-atom: (v from V(G)) do A(v) := 0;
F := []; FN := []; t := 1;
{"Поджог" стартового атома}
A(s) := t;   F := F + [s];
while ( F <> [] ) do
begin
   for v-atom: (v from F) do
          for u-atom: ( (u,v)  from  E(G) ) do
            if A(u) = 0 then
                 begin
                        A(u) := t+1;
                        FN := FN + [u];
```

```
                end;
   t := t + 1;   F := FN;  FN := []
end;
{ Конец Алгоритма Волна };
```

Утверждение 2.1. позволяет записать алгоритм вычисления матрицы расстояний **D**. В простейшем (не самом оптимальном варианте) требуется **n**-кратно применить волновой алгоритм к каждому атому графа.

Алгоритм Матрица_расстояний (G, D).

Вход: **G** - связный М-граф: **G = (V, E**);

Выход: **D** - **n*n** матрица расстояний графа, **n=|V|.**

Локальные данные: **A** - список числовых меток атомов.

```
for i:=1 to n do
begin
   Волна(G, v, A);
   for j:=1 to n do
    d[i,j] := A[j] -1;
end;
{Конец Алгоритма Матрица_расстояний}
```

На основе утверждения 2.1 могут быть вычислены также другие топологические и информационные индексы М-графов, например диаметр графа ***D(G)*** и индекс Винера ***W(G).***

2.4.2 Связность графа.

Рассмотрим алгоритмы, анализирующие связность графа.

Определение. Граф ***G*** - *связен*, если любая пара его вершин соединена простой цепью. Компонент связности - максимальный связный подграф графа ***G***. Несвязный граф имеет не менее двух компонент.

Алгоритм_функция Односвязен (G);

Вход: **G** - М-граф: **G = (V, E)**;
Выход: логическое значение функции:
Односвязен (G) = true, если граф связен;
Односвязен (G) = false, если граф несвязен;
Локальные данные: **A** - список числовых меток атомов.

```
{Инициализация} Односвязен:=true;
Волна (G, v1, A );
for j := 1 to n do
   if A[j]=0 then { Односвязен := false; выход; }
{Конец  Алгоритма Односвязен}
```

Утверждение 2.2. Пусть **s** - стартовый атом волнового алгоритма и проведена волновая разметка графа. Тогда для всех **v**-атомов с ненулевой **A**-меткой: **A(v) > 0** существует цепь, соединяющей **s**-атом с **v**-атомом, и следовательно, все эти атомы принадлежат одной связности графа.

Это утверждение позволяет записать алгоритм "вырезания" из несвязного графа его связной компоненты в следующем виде:

Алгоритм Связность (G, s, G1);

Вход: **G** - М-граф: **G = (V, E); s** - атом графа;

Выход: $\mathbf{G_1}$ - связный (с **s**-атомом) подграф **G**-графа.

Локальные данные: **A** - список числовых меток атомов.

```
{Инициализация} V(G1) := []; E(G1) := [];
if Односвязен(G) then begin G1 := G; выход end;
Волна (G, s, A );
for j := 1 to n do
   if A[j] > 0 then
   begin
          V(G1) := V(G1) + [vj];
          for u-атом: (u from Γ(v)) do
          E(G1) := E(G1) + [ (vj,u) ];
   end;
{ Конец Алгоритма Связность.}
```

Наиболее часто при работе с многосвязным графом стоит задача получения списка всех его связных компонент.

Алгоритм Связные_компоненты (G, {Gi});

Вход: **G** - М-граф: **G = (V, E);**

Выход: **Ф= { G_i }**- набор всех связных компонент графа **G**.

Локальные данные: **A** - список “обработанных” атомов.

G1 - граф - текущая связная компонента.

```
{Инициализация} Ф := {};
if Односвязен(G) then begin Ф:= {G}; выход end;
A:= [];
for v-атом: (v from V(G)) do
begin
   if not (v in A) then
   begin
         Связность(G,v,G1);
         Ф:= Ф + G1;
         for u-атом: (u from V(G1)) do
           A:= A + [u]
   end
end
{ Конец Алгоритма Связные_компоненты. }
```

2.4.3 Алгоритмы работы с кольцевыми системами М-графа.

Определение. Ребро **e=(u,v)** графа называется *кольцевым*, если при его удалении из графа связность графа не нарушается.

Обозначим через (**G \ (u,v)**) операцию вычеркивания ребра из **G**-графа, а через **B(e)** - числовую метку на e-ребре. Тогда алгоритм разметки ребер односвязного графа можно записать следующим образом:

Алгоритм Кольцевые_ребра(G, B);

Вход: **G** - М-граф: **G = (V, E);**

Выход: **B** - числовые метки на ребрах:

B(e)=1, если e-ребро кольцевое;

B(e)=0, если e-ребро некольцевое (ациклическое);

```
{Инициализация} for k := 1 to m do B[k] := 0;
if not Односвязен(G) then выход;
for ek-ребро: (ek from E(G)) do
   if Односвязен( G \ ek ) then B[k] := 1;
{Конец  Алгоритма Кольцевые_ребра}
```

Определение. Атом **v** является *кольцевым*, если хотя бы одно его ребро является кольцевым.

Для отметки кольцевых атомов введем специальную числовую метку ***R(v)***, которую будем называть "кольцевой степенью" ***v***-атома. Значение ***R(v)***

равно числу кольцевых ребер ***v***-атома, и оно может быть вычислено с помощью следующего алгоритма:

Алгоритм Кольцевые_атомы(G, R);

Вход: **G** - М-граф: **G = (V, E);**

Выход: **R** - числовые метки на атомах - кольцевая степень (**R(v)** - число кольцевых ребер v-атома).

```
for i := 1 to n do R[i] := 0;
if not Односвязен(G) then выход;
Кольцевые_ребра( G, B );
for vi-атом: (vi from V(G)) do
   for e-ребро: ( e=(vi,u), где u-атом: (u from Γ(vi)) ) do
         if B[e] = 1 then R[i] := R[i] + 1;
{Конец  Алгоритма Кольцевые_атомы}
```

На основе алгоритмов волновой разметки и определения кольцевых ребер М-графа был реализован алгоритм поиска колец в М-графе. Основная идея алгоритма заключается в следующем. Применим к какому-либо атому графа волновой алгоритм, и отметим атомы и ребра, на которых произошли “встречи” волн. Затем, проходя по графу обратно от мест “встречи волн” по уменьшению волнового числа ***A*(*v*),** получим список атомов, составляющих кольцо графа. Меняя начальный атом волновой разметки будем пытаться получить необходимое количество колец. Общее количество колец в односвязном графе есть его цикломатическое число ***C=m-n+1*** (где ***m*** -число ребер, ***n*** - число вершин).

Алгоритм Поиск минимальных колец графа (G,R)

Вход: **G** - М-граф: **G = (V, E);**

R(v) - кольцевые пометки на вершинах - кольцевая степень.

Выход: **RINGS** - список колец графа. **RINGS(i) = {v_1,v_2,...,v_k}** - список атомов, составляющих **i**-е кольцо графа.

Локальные данные: **USED** - список атомов.

```
RINGS:= [];    {Пустой список}
USED:= [];     {Список использованных для разметки
атомов                 графа}
C:= m-n+1;     {Цикломатическое число графа}
while |RINGS| < C do
begin
```

```
   MAX:= 0;
   for vi-атом: ( vi from G ) do
         if not (vi in USED) and (R(vi)>0) and
               (R(vi) > R(vMAX)) then MAX:= i;
   if MAX = 0 then выход; {не все кольца удалось найти}
   Поиск_Колец_С_Вершины(RINGS, VMAX);
   USED:= USED + [vMAX]
end;
{Конец алгоритма Поиск колец графа}
```

Алгоритм Поиск_колец_с_вершины (G, RINGS, Start)

Вход: **G** - М-граф: **G = (V, E);**

RINGS - список найденных колец графа

Start - начальный атом для разметки

Выход: **RINGS** - возможно пополненный список колец графа.

Локальные данные: **MEETS** - список “встреч” волны.

A - список пометок атомов

Замечание: при работе будем использовать несколько видоизмененный алгоритм волновой разметки **Волна_И_Встречи (G, s, A,MEETS),** который возвращает список “встреч” волны **MEETS**. **MEETS(i) = v**, если “встреча” произошла на атоме; **MEETS(i) = e**, если - на ребре; **Pow(MEETS(i)) -** количество “встреч”, происшедших на данной вершине или ребре. Примеры “встреч” на ребре и на вершине приведены на рисунке 2.3

Алгоритм ВОЛНА_И_ВСТРЕЧИ(G, Start, A, MEETS);

```
for M: (M from MEETS) do
   Поиск_Колец_С_Места_Встречи(G,RINGNS,M)
{Конец алгоритма Поиск колец с вершины}
```

Алгоритм Поиск колец с “места встречи” (G, RINGS, M)

Вход: **G** - М-граф: **G = (V, E);**

RINGS - список колец графа M - место встречи (**v**-атом или **e**-ребро)

Выход: **RINGS** - возможно пополненный список колец графа.

Локальные данные: **A,B,D** - список пометок атомов

R - список атомов кольца

```
B(G):=-1; T:=1; B[M]:= T; Pw:= Pow(M);
while (Pw > 0) do
begin
   D(G):= 0;
   repeat
         for v-атот: (v from B) do
         for u-атом: ((u,v) from E(G)) and (A(v)>=A(u) do
         begin
               if (B(u)=T) and (M<>(v,u)) then
                 begin Pw:= Pw-1; D(V):= 1; D(u):= 1; goto L1
end
               else if (B(u)=T+1) then
                 begin       Pw:= Pw-1; D(u):= D(u)-1; goto L1
end
               else if B(u)=-1 then B(u):= T+1
         end
   until T > |V|+2
L1:
   Max:= max(B(v): B(v)>-1);
   N:= Count(v: B(v)>-1 and B(v)=Max and D(v)=0);
   if N > 0 then {Найдено кольцо из N атомов}
   begin
         R:= (v: B(v)>-1 and B(v)=Max and D(v)=0);
         if Правильное_Кольцо(R) and (R∉RINGS)
               then RINGS:= RINGS + [R]
   end;
end;
{Конец алгоритма Поиск колец с места встречи}
```

Алгоритм_функция Правильное кольцо (G, R)

Вход: G - M-граф: G = (V, E); R - список атомов кольца

Выход: True, если кольцо можно использовать; False - иначе. R - “исправленный” вариант кольца.

Локальные данные: A - список пометок атомов

Замечание Кольцо не должно содержать менее 3-х атомов, каждый атом из кольца должен содержать двух соседей из него же.

```
for v-атом: (v from R) do
   if Count(u: ((u,v) from E(G)) and (u from R)) < 2
         then R:= R - [v];
for v-атом (v from R) do
```

```
   if Count(u: ((u,v) from E(G)) and (u from R)) > 2
         then begin RESULT:= False; выход end;
RESULT:= True
{Конец алгоритма Правильное кольцо}
```

Определение. Будем называть два кольца R_1 и R_2 графа **G** *смежными*, если у них есть общее ребро, т.е. $R_1 \circ R_2$, если $\exists e \in E(G)$: . $e \in R_1$ и $e \in R_2$. На основе графа **G** с найденными кольцами определим *кольцевой граф* **GR**. Вершины этого графа соответствуют кольцам графа **G**. Две вершины v_i, v_j графа **GR** связаны (т.е. существует ребро $e=(v_i,v_j)$), если соответствующие кольца графа **G** - смежные $(R_i \circ R_j)$. Отдельные связные компоненты графа **GR** будут соответствовать кольцевым системам графа **G**.

Алгоритм Построение кольцевого графа (G, RINGS, GR)

Вход: **G** - М-граф: **G = (V, E);**

RINGS - список колец графа

Выход: **GR** - соответствующий кольцевой граф

Локальные данные: **A** - список пометок атомов

```
E(GR):= [];
V(GR):= [];
for R-кольцо: (R from RINGS) do
   V(GR):= V(GR) + [V с пометкой R];
for v-вершина (v from V(GR)) do
   for u-вершина (u from V(GR)) and (u <> v) do
         if Rv°Ru then E(GR):= E(GR) + [(v,u)]
{Конец алгоритма Построение кольцевого графа}
```

На основании списка колец М-графа и списка кольцевых систем можно решить практически любую задачу по определению химического типа кольца или кольцевой системы. Рассмотрим алгоритм определения ароматичности колец М-графа. Основная идея алгоритма заключается в проверке каждого кольца графа на соответствие правилу Хюккеля [47]: количество свободных π-электронов атомов ароматического кольца ***(N-2)/4>0*** и ***(N-2)*** должно делиться без остатка на **4.**

Алгоритм Пометка ароматических колец (G, RINGS, AROMA)

Вход: **G** - М-граф: **G = (V, E);**

RINGS - список колец графа

Выход: **AROMA** - список ароматических колец.

Локальные данные: **AR_AT** - текущий список ароматических атомов.

```
AROMA:= [ ]; AR_AT:= [ ];
repeat
   WasAroma:= False;
   for R-кольцо: (R from RINGS) do
         if Ароматическое_Кольцо(G,R,AR_AT) then
         begin
               AROMA:= AROMA + [R];
               for v-атом: (v from R) do
                     AR_AT:= AR_AT + [v];
               WasAroma:= True
         end
until not WasAroma; {Конец алгоритма Пометка
ароматических колец }
```

Алгоритм_функция Ароматическое кольцо (G, R, AR_AT)

Вход: **G** - М-граф: **G = (V, E);**

R - список атомов кольца

AR_AT - текущий список ароматических атомов.

Выход: **True** - если кольцо ароматическое, **False** - иначе.

Локальные данные: **Pi** - список пометок на атомах: **Pi(v)=**количество свободных π-электронов на атоме **v**.

```
ОпределитьПиЭлектроны(G,R,Pi);
N:= 0;
for v-атом: (v from R) do
   if Pi(v) < 0 then
         begin RESULT:= False; выход end
   else N:= N + Pi(v);
RESULT:= ((N-2) div 4 > 0) and ((N-2) mod 4 = 0);
{Конец алгоритма Ароматическое кольцо}
```

В алгоритме использованы операции **mod** - остаток от целочисленного-деления и **div** - целочисленное деление.

Алгоритм Определить π-электроны (G, R, AR_AT)

Вход: **G** - М-граф: **G = (V, E);**

R - список атомов кольца

AR_AT - текущий список ароматических атомов.

Выход: **Pi** - список пометок на атомах кольца **R**: **Pi(v)=**количество свободных π-электронов на атоме **v**.

```
for I:= 1 to n do Pi(i):= 0;
for v-атом: (v from R) do
begin
   Val:= Валентность(V);
   RS:= 0; RD:= 0; C:= 0;
   for u-атом ( (u,v) from E(G)) do
   begin
          c:= c+1;
          if B((u,v)) = "двойная кольцевая" then
          begin
          if (u∈R) or (u∈AR_AT) then RD:= RD+1
          end
          else B((u,v)) = "одинарная кольцевая"
                then RS:= RS+1;
   end;

   if (RS >= 1) and (RD = 1) then Pi[v]:= 1
   else if (RS = 2) and (Val = 2) and
          (A(v) from ['O', 'S', 'SE', 'TE'])
   then PiBonds[N]:= 2
   else if (RS >= 2) and (Val = 3) and
          (A(v) from ['N', 'P', 'AS'])
   then PiBonds[N]:= 2
   else PiBonds[N]:= -1;
end;
{Конец алгоритма Определить π-электроны }
```

Валентность атома определяется как сумма кратностей связей + число водородов, присоединенных к атому, - заряд на атоме.

Надежно работающий алгоритм распознавания ароматики, описанный выше, позволил решить такую непростую проблему, как установление изоморфизма ароматических М-графов. На рисунке 2.6 приведен пример двух М-графов, эквивалентных с химической точки зрения, но различных по топологии (по разному помечены ребра). Введя новый тип пометки - "ароматическое" ребро, удалось совместить топологию подобных графов.

При генерации графов требуется не просто сравнить два М-графа, а найти М-граф изоморфный данному в списке М-графов (например, провести поиск в химической базе данных). В этом случае будем применяется инвариант, являющийся упорядоченным списком локальных инвариантов вершин, - "первичный ключ" поиска в СБД, значения которого на изоморфных М-графах совпадают. В качестве такого первичного ключа, например, можно

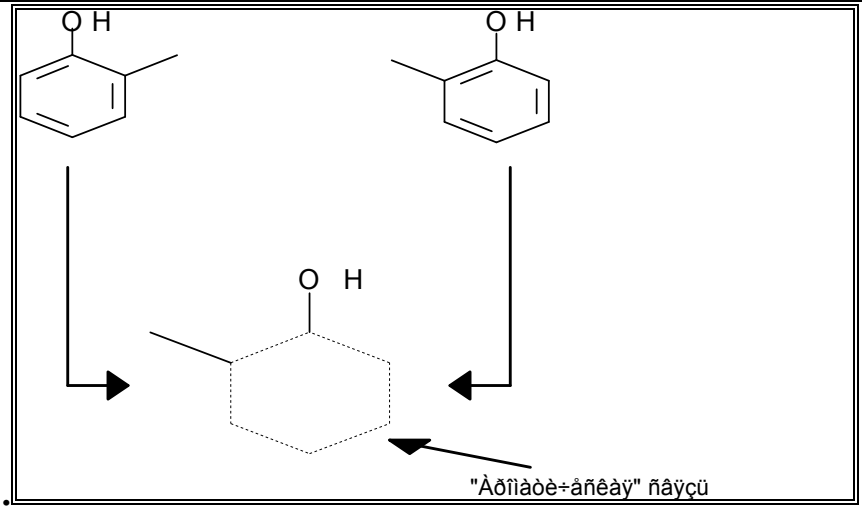


Рисунок 2.6. "Ароматическое" преобразование М-графов с целью установления изоморфизма.

использовать символьную строка вида "$\mathbf{M_1n_1M_2n_2...M_kn_k}$", где $\mathbf{M_k}$ - символьная метка атома, построенная на основе использования маркеров, как описано выше, $\mathbf{n_k}$ - число атомов в М-графе, метки которых совпадают и равны $\mathbf{M_k}$. Внутри строки-ключа метки отсортированы в лексикографическом порядке. Поскольку на строке определены отношения больше/меньше/равно, то возможно проведение очень быстрого поиска в списке М-графов, совпадающих по первичному ключу с искомым.

2.5 Выводы.

Описана методическая и алгоритмическая основа для проведения описания молекул на основе дескриптивного подхода, разработанного для распознавании изображений - структурных объектов, имеющих много общих черт с молекулярными графами. Отмечено, что существует глубокая аналогия между распознаванием (классификацией) молекулярных структур и распознаванием изображений, которая основана на незаданности стандартного описания исследуемых объектов. Приведена краткая характеристика алгебраического подхода к распознаванию, разработанного Ю.И.Журавлевым. Рассмотрены различные способы выбора и классификации примитивов описания, на основе которых путем индуктивного порождения формируются структурные спектры молекул заданной степени сложности. Доказана теорема о расщеплении столбцов матрицы "граф-признак". Проанализированы свойства матриц "молекула-дескриптов", формируемых по маркированным графам обучающей выборки. Показано, что выбор локального свойства атомов, с помощью которого формируется маркерная разметка вершин графа, является определяющим моментом для последующего индуктивного построения признаковых пространств, среди которых будет проводится селекция пространства, адекватного для анализируемого свойства молекул. Описан минимальный базовый набор алгоритмов для работы с молекулярными графами, который служит основой для программной реализации различных более сложных алгоритмов анализа и преобразования М-графов.

Глава 3. Применение эволюционных алгоритмов для решения QSAR-задач

3.1 Метод Группового Учета Аргументов (МГУА)

Результатом построения структурного спектра обучающей СБД (в заданных ограничениях) является (**N*****M**)-матрица “молекула-фрагмент” ***X*** (где **N** - количество структур, **M** - количество структурных дескрипторов). При этом возникает проблема "информационного взрыва": формируются “*очень широкая*” ***(N<<M)*** матрица ***X***, обработка которой традиционными методами анализа данных в рамках того или иного программного пакета вызывает затруднения, поскольку уже при небольших значениях ***N=50-100*** и относительно несложных **k-**фрагментах (**k=4-6** атомов) могут получаться очень “длинные” структурные спектры: ***M=2000-5000***.

Идея применения полного фрагментного перечисления СБД для построения QSAR-зависимостей имеет давнюю историю [315], однако в практическом плане ее программное воплощение наталкивалось на две трудности:

1. технические ограничения оперативной и долговременной памяти компьютеров;
2. отсутствие в стандартных пакетах анализа данных модулей для обработки “очень широких” таблиц **X** “объект-признак” при построении тех или иных зависимостей.

Практическое использование “структурных спектров” СБД является весьма привлекательным по ряду важных причин:

1. “Структурный спектр”, сформированный в заданных ограничениях на сложность перечисляемых фрагментов, содержит подробное и избыточное описание молекул СБД, т.е. содержит достаточную информацию для формирования QSAR-моделей.
2. Известно, что с теоретической точки зрения “структурный спектр” является основой для формирования произвольных инвариантов молекулярных графов [4, 5].
3. Структурные фрагменты, как признаки молекулы, имеют ясную и четкую содержательную интерпретацию для химиков-практиков, что

является хорошей основой для проведения возможных обобщений QSAR-моделей и формулирования на их основе качественных выводов об особенностях проявления исследуемого свойства.

В главе излагается метод самоорганизации моделей, предложенный Ивахненко А.Г. с соавторами [316,317], (метод группового учета аргументов - МГУА), который модифицирован нами для проведения QSAR-анализа структурных спектров. Использование МГУА позволяет не только отбирать значимые переменные в ходе построения QSAR-уравнений, но также дает возможность выполнять функциональные преобразования дескрипторов в ходе расчетов.

3.1.1 Общая схема МГУА

Пусть задана матрица ***X (NxM)*** "объект-признак". В заданном классе функций необходимо построить функцию $F(X_{j1},X_{j2},...,X_{jk})$ от **K** столбцов-переменных, где переменные $X_1,X_2,...,X_k$ отбираются из большого набора дескрипторов $X_1,X_2,...,X_M$ (т.е. среди столбцов матрицы ***X,*** **M>> k**). При построении функции $F(X_{j1},X_{j2},...,X_{jk})$ необходимо минимизировать заданный критерий ***Ф,*** например, критерий наименьших квадратов невязки:

$$E = (C - F(X_1,X_2,...,X_k)); \qquad \Phi: (E,E) \rightarrow min \qquad (3.1)$$

Построение функции $F(X_{j1},X_{j2},...,X_{jk})$ проводится по шагам, которые называются селекциями.

Первая селекция. Построим (переберем) ***все уравнения от двух переменных*** : $\overline{S}(1) = F(\overline{X}_{j1}, \overline{X}_{j2})$ и упорядочим их по возрастанию критерия ***Ф***:

$$\Phi_1(1) \le \Phi_2(1) \le ... \le \Phi_L(1)$$

Если функция ***F*** симметрична (т.е. $F(\overline{A},\overline{B}) = F(\overline{B},\overline{A})$), то число **L** таких уравнений равно $L = \frac{M \cdot (M-1)}{2}$.

Выберем первые Q_1 лучших уравнений, которые примут участие во *второй селекции*: $\{(S_p(1),\Phi_p\}$, $p=1,...,Q_1$. Проведем замену переменных и сформируем новые Q_1 вектор-столбцов $\overline{S}_p$: $\overline{S}_p(1) = F_p(\overline{X}_{j1}, \overline{X}_{j2})$, $p=1,...,Q_1$, где ***p*** - номер уравнения.

(k+1)*-ая селекция.** Пусть на ***k-й селекции получено Q_k столбцов-признаков: $\{\overline{S}(k)_p\}$ $p=1,...,Q_k$. Построим (Q_k*M) уравнений от двух переменных: $\overline{S}(k+1) = F(\overline{X}_j, \overline{S}(k)_p)$; $j=1,...,M;\ p=1,...,Q_k,$ и порядочим их по возрастанию критерия Φ. Отберем Q_{k+1} первых лучших уравнений:

$$\{(S_p(k+1), \Phi_p(k+1)\},\quad p=1,...,Q_{k+1};\ \Phi_p(k+1) \le \Phi_p(k+1) \le ... \le \Phi_P(k+1)$$

Заменой переменных сформируем набор новых вектор-столбцов для исследования в следующей селекции:

$\overline{S}_q(k+1) = F_q(\overline{X}_j, \overline{S}(k)_q)$, $q=1,...,Q_{k+1}$.

Условие остановки. Селекции останавливаются при достижении заданного **K**. Кроме этого, для остановки может быть использовано условие неубывания критерия Φ: $|\Phi_1(i)-\Phi_1(i-1)|<\varepsilon,$ где $\Phi_1(i)$ - минимальная величина критерия Φ для ***i***-й селекции.

Формирование искомой функции. “Раскроем скобки” в уравнениях ***(k-1)***-й селекции $\{S(k-1)_p\}$ и перейдем к исходным переменным:

$$S(k-1)_1 = F(\overline{X}_{j_1}, \overline{S}(k-2)_{q_1}) = F(\overline{X}_{j_1}, F(\overline{X}_{j_2}, S(k-3)_{q_2})) = ... =$$

$$F(\overline{X}_{j_1}, F(\overline{X}_{j_2}, F(..., \overline{X}_{j_k}))) = F_1^*(\overline{X}_{j_1}, \overline{X}_{j_2}, ..., \overline{X}_{j_k});$$

.....

$S(k-1)_p = F_p^*(\overline{X}_{j_1}, \overline{X}_{j_2}, ..., \overline{X}_{j_k})$; $p=1,...,Q_{k-1}$.

Таким образом, получим ***семейство Q_{k-1} различных уравнений*** от ***K*** переменных-столбцов, качество которых (критерий **Ф**) изменяется в пределах $(\Phi_1(k-1), \Phi_{Qk-1}(k-1))$.

Параметрами метода (при всех прочих равных условиях) является последовательность натуральных чисел $Q_n=\{Q_1,Q_2,...,Q_k\}$, где Q_p определяет число отбираемых уравнений ***p***-ой селекции. Если $Q_n=\{1,1,...1\}$, то получаем аналог “пошагового построения” заданной функции ***F***.

МГУА схема для линейной функции. Пусть необходимо провести QSAR-моделирование свойства **C** и построить линейную функцию от **k** переменных:

$$F(X_1,X_2,...,X_k) = b_0 + b_1*X_1 + b_2*X_2 + ... + b_k*X_k;$$

$$E = (C - F(X_1,X_2,...,X_k)); \quad \Omega: (E,E) \rightarrow \min.$$

где переменные $X_1,X_2,...,X_k$ отбираются на основе критерия Ω из большого набора дескрипторов $X_1,X_2,...,X_M$ (т.е. среди столбцов QSAR матрицы ***X*** "молекула-фрагмент"; **M>> k**). Будем строить модель **F** шаг за шагом (которые назовем *селекциями*) следующим образом.

Первая селекция. Система строит всевозможные регрессионные уравнения с **двумя** переменными:

$$S[1] = F(X_i, X_j) = b_0 + b_1*X_i + b_2*X_j ; \quad 1 \le i < j \le M.$$

где X_i и X_j- столбцы QSAR-матрицы. Общее число таких уравнений равно $Z_1=M*(M-1)/2$. Согласно оптимизационному критерию Ω, система отбирает налучшие Q_1 уравнений {**S[1]**} , которые будут принимать участие во втором селекции $(Q_1<<Z_1)$.

Результатом первой селекции являются:

1. Набор вектор-столбцов $\{S[1]_q, q=1,...,Q_1\}$ и
2. Список записей параметров для каждого **q**-ого столбца: $(q: i, j, b_0, b_1, b_2)$.

Вторая селекция. Система добавляет переменные X_i к уравнениям, отобранным на первом шаге, используя линейную функцию от **двух** переменных:

$$S[2] = F(X_i, S[1]_q) = b_0 + b_1*X_i + b_2*S[1]_q ; \quad i = 1,...,M; \quad q = 1,...,Q_1$$

где X_i - **i**-ый столбец QSAR-матрицы и $S[1]_q$ - **q**-ый столбец, отобранный на первой селекции. Общее число построенных уравнений равно $Z_2=M*Q_1$. Согласно критерию Ω, система отбирает налучшие Q_2 уравнения {**S[2]**} $(Q_2<<Z_2)$. Результом второй селекции являются:

1.Набор вектор-столбцов $\{ S[2]_p, \quad p=1,...,Q_2 \}$ и

2.Список записей параметров для каждого **p**-ого столбца: $(p: i, q, b_0, b_1, b_2)$.

(k+1)-ая селекция. Система строит уравнения, используя **(i,q)-пары** переменных:

$$S[k+1]=F(X_i,S[k]_q)=b_0+b_1*X_i+b_2*S[k]_q; \quad i= 1,...,M; \quad q=1,...,Q_k.$$

где X_i - **i**-ый столбец QSAR-матрицы и $S[k]_q$ - **q**-ый столбец, отобранный на **k**-ой селекции. Общее число построенных уравнений равно $Z_{k1}=M*Q_k$. Согласно критерию Ω, система отбирает налучшие Q_{k+1} уравнений {**S[k+1]**}.

Критерием остановки является проведение заданного числа селекций ***K***.

Если “открыть скобки” в уравнениях, полученных на **k**-ой селекции, и подставить в них исходные переменные $\mathbf{X_j}$, то получим семейство целевых уравнений: $\mathbf{(X_1,X_2,...,X_k)_q=(b_0+b_1*X_1+b_2*X_2+...+b_k*X_k)_q;\quad q= 1,...,Qk.}$

3.1.2 Модифицированная схема построения семейств QSAR-уравнений

Основной постулат МГУА схемы для оптимизации линейных уравнений состоит в том, что для задач прогнозирования целесообразно несколько снизить точность определения оценок коэффициентов уравнений регрессии, но за счет этого придать ему большую “прогнозирующую силу”. Это достигается за счет использования *критериев селекции специального вида*: **критерия несмещенности и критерия регулярности**.

Для QSAR-задачи решение ищется в классе линейных функций:

$$\overline{F}(\overline{X}_1,...,\overline{X}_K) = b_0 + \sum_{j=1}^{K} b_j \cdot \overline{X}_j = \hat{y}; \qquad \overline{y} = \overline{F} + \overline{\varepsilon} = \hat{y} + \overline{\varepsilon};$$

Критерий ***Φ***: $\Phi = (\overline{\varepsilon},\overline{\varepsilon}) = \sum_{i=1}^{N} \varepsilon_i^2$;

Модификация общей схемы МГУА связана с преодолением *возможной высокой взаимной корреляции столбцов* $\boldsymbol{X_j}$ в матрице ***X***, которая может привести к сильной взаимной корреляции столбцов $\boldsymbol{\{S_p(k)\}}$, передаваемых на ***(k+1)***-ю селекцию. *Будем отбирать столбцы-признаки* ***S(k)*** *так, чтобы их попарная корреляция не превосходила некоторый порог* $\boldsymbol{R_T\ (0 \le R_T \le 1)}$.

Пусть задан порог $\boldsymbol{R_T(k)}$. Тогда имеем следующую схему ***(k+1)***-й селекции.

(k+1)-я селекция QSAR. Пусть на ***k***-й селекции получено $\boldsymbol{Q_k}$ столбцов-признаков: $\{\overline{S}(k)_p\}$, $\boldsymbol{p=1,...,Q_k;}$ $\boldsymbol{R_{pq}(k)<R_T(k);\ 1 \le p<q \le Q_k;}$

Построим $\boldsymbol{(Q_k*M)}$ уравнений:

$$\overline{y} = \overline{F}(\overline{X}_j,\overline{S}(k)_p) + \overline{\varepsilon} = b_0 + b_1 \cdot \overline{X}_j + b_2 \cdot \overline{S}(k)_p + \overline{\varepsilon}; \quad \Phi = (\overline{\varepsilon},\overline{\varepsilon}) = \sum_{i=1}^{N} \varepsilon_i^2 \to \min;$$

и упорядочим их по возрастанию ***Φ***:

$\overline{F}_1, \overline{F}_2, \overline{F}_3, ..., \overline{F}_L$; $\boldsymbol{L=Q_k*M;}$ $\boldsymbol{\Phi_1 \le \Phi_2 \le \Phi_3 \le ... \le \Phi_L.}$

Обозначим через $\boldsymbol{R_{pq}}$ коэффициент корреляции векторов-функций $\overline{F}_p$ и $\overline{F}_q$:

$$R_{pq}=Corr(\overline{F}_p, \overline{F}_q)$$

Формирование столбцов-признаков, передаваемых на следующую селекцию, проведем следующим образом:

$$\overline{S}_1(k+1) = F_1(\overline{X}_j, \overline{S}(k)_{p_1});$$

$$\overline{S}_2(k+1) = F_{q_2}(\overline{X}_j, \overline{S}(k)_{p_2});\ \text{где}\ q_2 = \min_p\{p: R_{lp} < R_T(k)\}$$

$$\overline{S}_3(k+1) = F_{q_3}(\overline{X}_j, \overline{S}(k)_{p_3});\ \text{где}\ q_3 = \min_p\{p: R_{lp} < R_T(k), R_{q_2 p} < R_T(k)\}$$

.....

Таким образом, второй вектор-признак ***$S_2(K+1)$*** формируется путем поиска первого уравнения F_{q_2}, взаимная корреляция которого с первым столбцом не превышает порог ***$R_T(k)$***; третий столбец должен иметь коэффициент парной корреляции с первым и вторым столбцами не выше ***$R_T(k)$*** и так далее. Отобранные ***Q_{k+1}*** вектор-столбцов ***$\{S_p(k+1)\}_1^Q$*** передаются на следующую селекцию.

Динамическое функциональное преобразование признаков. В ходе построения QSAR-уравнений проводится следующая модификацию МГУА алгоритма. На первой селекции могут быть сформированы и отобраны новые признаки, полученные в результате преобразования исходных столбцов матрицы **X** “объект-признак”: $\mathbf{Z_j = f(X_j)}$. Система оптимизирует регрессионную функцию **F** только от одной переменной. Реализованный набор функций одного переменного $\mathbf{\{f_1, f_2,..., f_m\}}$ позволяет вычислять следующие новые дескрипторы $\mathbf{\{ Z_p: }$ ***$p=1, ..., m$***$\mathbf{\}}$:

$\mathbf{X_j}$: $\quad \mathbf{C = (b_0 + b_1*X_j) + E_j = S[1]_j + E_j}$; $\quad$ ***$j = 1,...,M$***

$\mathbf{Z_1 = f_1(X_j)}$: $\quad \mathbf{C = (b_0 + b_1*Z_1) + E_{1j} = S[1]_{j1} + E_{1j}}$; $\quad$ ***$j = 1,...,M$***

$\mathbf{Z_2 = f_2(X_j)}$: $\quad \mathbf{C = (b_0 + b_1*Z_2) + E_{2j} = S[1]_{j2} + E_{2j}}$; $\quad$ ***$j = 1,...,M$***

. . . . . .

$\mathbf{Z_m = f_m(X_j)}$: $\quad \mathbf{C = (b_0 + b_1*Z_q) + E_{mj} = p[1]_m + E_{mj}}$; $\quad$ ***$j = 1,...,M$***

где **C** - вектор-столбец заданного свойства.

Каждый столбец $\mathbf{S[1]_j, S[2]_{j1}, S[1]_{j2}, ..., S[1]_{jm}}$ *независимо участвует* в формировании списка налучших $\mathbf{Q_1}$ уравнений на основе критерия селекции

Ω. Дополнительно существует еще один критерий селекции: взаимная корреляция между любыми двумя отбираемыми столбцами не должна превышать определенный порог **Tr,** заданный экспертом **(**например, **Tr= 0.93-0.97)**. Вторая, третья, четвертая и последующие селекции проводятся аналогично по схеме, описанной выше.

Рассмотрим используемые в МГУА критерии оптимизации ***Ф***.

Критерий регулярности. В качестве критерия селекции используется величина среднеквадратической ошибки, вычисленная на тестовой выборке размера N_T.

$$\Delta^2 = \frac{\sum_{i=1}^{N_T}(\varphi_i - \hat{\varphi}_i)^2}{\sum_{i=1}^{N_T}(\varphi_i - \overline{\varphi})^2} 100\%, \ \hat{\varphi}_i = F(X_1, X_2, ..., X_K);$$

гдеΔ- среднеквадратическая ошибка на проверочной последовательности, %

$\hat{\varphi}_i$ - значение прогноза в ***i***-й точке, ***i=1,...,N***;

φ_i - действительное (экспериментальное) значение в той же точке;

$\overline{\varphi}$ - среднее φ.

Критерий несмещенности. Для расчета критерия несмещенности экспериментальные данные обучающей выборки ранжируются и делятся на две части. Объекты с четными номерами образуют первую последовательность R_1, а с нечетными - вторую последовательность данных R_2.

На каждой селекции синтез уравнений выполняется два раза. Сначала первая последовательность данных является обучающей, а вторая - проверочной: $|R_1|=N_L$, $|R_2|=N_T$ ($N_L+N_T=N$). Вектор оценки “свойства” по всей выборке обозначим Y^*. Затем, ***R1*** и ***R2*** меняются местами: первая служит проверочной, а вторая - обучающей: $|R_1|=N_T$, $|R_2|=N_L$. Новый вектор оценки “свойства” обозначим Y^{**}.

Каждое из найденных уравнений регрессии оценивается по величине δ среднеквадратического отклонения, рассчитанного по всем ***N*** точкам обоих

последовательностей: $\delta = \frac{1}{N} \cdot \sum_{r=1}^{N} \left(y_r^* - y_r^{**} \right)^2$. Из всех уравнений регрессии на первом ряду выбираются **Q** уравнений, имеющих меньшую оценку δ.

Критерий несмещенности решений для всей селекции Δ определяется как среднее значение показателя несмещенности для ***Q*** отобранных уравнений:

$$\Delta = \frac{1}{Q} \sum_{i=1}^{Q} \delta_i$$

Правило остановки следующее: ряды селекции наращиваются до тех пор, пока критерий несмещенности решений падает.

Модифицированный критерий несмещенности. В качестве критерия оптимизации рассмотрим некоторую модификацию критерия несмещенности, ориентированную на получение максимально прогностичных моделей.

Как и в случае критерия несмещенности, все имеющиеся экспериментальные данные упорядочиваются по величине экспериментального свойства и делятся на две части: объекты с четными номерами образуют первую последовательность R_1, а с нечетными - вторую последовательность данных R_2. На каждой селекции синтез уравнений выполняется два раза. Сначала первая последовательность данных является обучающей, а вторая - проверочной: $/R_1/=N_L$, $/R_2/=N_T$. Уравнения регрессии, полученные при этом, обозначим $Y^*=b\hat{}_0+b\hat{}_1x_1+b\hat{}_2x_2+...+b\hat{}_nx_n$. Затем, наоборот, первая служит проверочной, а вторая - обучающей. $/R_1/=N_T$, $/R2/=N_L$. Уравнения регрессии, полученные при этом, обозначим $Y^{**}=b\hat{}\hat{}_0+b\hat{}\hat{}_1x_1+b\hat{}\hat{}_2x_2+...+b\hat{}\hat{}_nx_n$. Пусть Y^*_i - оценка свойства для ***i***-го соединения обучающей выборки ***TS***, полученное для <u>тестовой</u> последовательности (половина таких значений рассчитана по уравнению для Y^*, вторая - для Y^{**}). В качестве значения критерия оптимизации ***Ф*** примем величину квадрата коэффициента множественной корреляции $R^2 = 1 - \frac{\sum (v_i - Y_i^*)^2}{\sum (v_i - \bar{v})^2}$, где v_i - экспериментальное значение свойства, $\bar{v}$ - среднее значение свойства по СБД. "Итоговые" весовые коэффициенты b_i определим как $\frac{b_i\hat{} + b_i\hat{}\hat{}}{2}$.

Достоинство так введенного нами критерия ***Φ*** состоит в том, что в результате мы получаем максимально прогностичные уравнения, для которых все оценки статистических параметров (стандартное отклонение, регрессионный коэффициент корреляции, критерий Фишера [318]) близки к таким же оценкам, полученным в результате “скользящего контроля” - метода, применяемого для оценки прогностичности модели после завершения ее построения. Здесь такая проверка осуществляется “на ходу”, по мере наращивания сложности моделей: наиболее прогностичные модели отбираются по ходу построения.

3.2 МГУА-поиск сложных фрагментов, адекватных свойству

3.2.1 Обобщенные фрагменты М-графов.

Определение. Пусть задано разбиение структурного спектра на подмножества фрагментов: $P = \{p_1, p_2, ..., p_q\} = p_1 \oplus p_2 \oplus ... \oplus p_q$. Каждое такое подмножество фрагментов $\boldsymbol{p_i}$ назовем *обобщенным фрагментом*. Обобщенный фрагмент $\boldsymbol{p_i}$ задается множеством фрагментов его составляющих: $p_i = \{f_1^i, f_2^i, ..., f_k^i\} = \bigcup_{j=1}^{k_i} f_j^i$

Утверждение 3.1 Пусть задан обобщенный фрагмент $p=\{f_1,f_2,...,f_k\}$: $p = \bigcup_{j=1}^{k} f_j$. Тогда вектор-столбец $\overline{X}[p]$, соответствующий числу повторений обобщенного фрагмента *p* в обучающей выборке, равен сумме соответствующих столбцов матрицы *X*: $\overline{X}[p] = \sum_{j=1}^{k} \overline{X}_j = \sum_{j=1}^{k} \overline{X}[f_j]$

Утверждение 3.2 Выберем в QSAR-матрице ***X*** набор столбцов с номерами $(j_1,j_2,...,j_k)$ и суммируем их: $\overline{X}_z = \sum_{l=1}^{k} \overline{X}_{j_l}$. Тогда вектор-столбец $\overline{X}_z$ будет указывать число повторений “обобщенного” фрагмента *p* в обучающей выборке, где обобщение задается формулой $p = \bigcup_{l=1}^{k} f_{j_l}$.

Справедливость утверждений 3.1 и 3.2 вытекает из построения матрицы ***X***, ***j***-й столбец которой указывает число повторений фрагмента $\boldsymbol{f_j}$ в обучающей

СБД. Утверждение 3.2 позволяет построить МГУА-схему автоматического поиска обобщений фрагментов, адаптированных для заданного свойства ***Y*** М-графов.

1-я селекция построения “обобщения фрагментов”.

По матрице ***X*** построим ***M*(M-1)/2*** функций:

$$\bar{y} = F(X_{j_1}, X_{j_2}) + \bar{\varepsilon} = b_0 + b_1 \cdot (\bar{X}_{j_1} + \bar{X}_{j_2}) + \bar{\varepsilon}\,,\ \Phi = (\bar{\varepsilon}, \bar{\varepsilon}) = \sum_{i=1}^{N} \varepsilon_i^2 \to \min\,,$$

которые упорядочим по возрастанию критерия ***Ф***. Отберем лучшие Q_1 уравнений с учетом критерия максимальной попарной коррелированности векторов с порогом $R_T(1)$. На основе отобранных уравнений сформируем новые столбцы признаки:

$$\bar{S}_l(1) = \bar{X}_{j_1}(l) + \bar{X}_{j_2}(l)\,,\ l=1,\ \dots\ ,Q_1$$

где $\bar{X}_{j_1}(l)$ и $\bar{X}_{j_2}(l)$ - вектор-столбцы, вошедшие в ***l***-ое отобранное уравнение.

(k+1)-я селекция “обобщения фрагментов”.

Пусть на ***k***-й селекции получено Q_k столбцов-признаков:

$$\{\bar{S}(k)_l\}\,,\ l=1,\dots,Q_k;\ Corr(\bar{S}(k)_l, \bar{S}(k)_q) < R_T(k)\,,\ 1 \le l < q \le Q_k.$$

Построим $(Q_k * M)$ уравнений:

$$\bar{y} = \bar{F}(\bar{X}_j, S(k)_l) + \bar{\varepsilon} = b_0 + b_1 \cdot (\bar{X}_j + S(k)_l) + \bar{\varepsilon}\,,\ \Phi = (\bar{\varepsilon}, \bar{\varepsilon}) \to \min$$

Упорядочим уравнения по возрастанию критерия ***Ф*** и отберем Q_{k+1} первых уравнений с учетом критерия максимальной попарной коррелированности с порогом $R_T(k+1)$. На основе отобранных уравнений сформируем столбцы-признаки:

$$\bar{S}(k+1)_l = \bar{X}_{j_1}(l) + \bar{S}(k)_{q_1}\,,\ l=1,\dots,Q_{k+1},$$

где $\bar{X}_{j_1}(l)$ и $\bar{S}(k)_{q_1}$ - вектор-столбцы, вошедшие в *l*-е отобранное уравнение.

Критерием остановки является проведение заданного числа селекций ***K***. “Раскроем скобки” для столбцов-признаков $\{S(k)_l\}$, $l=1,\dots,Q_k$, полученных на ***k***-ой селекции. Для фиксированного ***l*** имеем:

$$\bar{X}^*(l) = \bar{S}(k)_q = \bar{X}_{j_1}(l) + \bar{S}(k-1)_{q_1} = \dots \qquad \textbf{(3.2)}$$

$$= \bar{X}_{j_1} + \bar{X}_{j_2} + \bar{S}(k-2)_{q_2} = \dots = \bar{X}_{j_1} + \bar{X}_{j_2} + \dots + \bar{X}_{j_{k+1}}\,,\ l=1,\dots,Q_k$$

Согласно утверждению 3.2, сумма (3.1) соответствует обобщенному фрагменту $p = \bigcup_{q=1}^{k+1} f_{jk}$. Согласно МГУА-построению "обобщенный фрагмент" ***p*** наиболее адекватен (для заданного исходного списка фрагментов) аппроксимации вектора-свойства **Y**. Для каждого ***l (l=1,...,Q_k)*** будет получено отдельное "обобщение" ***p_l***, которым будут соответствовать вектора-столбцы

$$\overline{X}^*(1), \overline{X}^*(2), ..., \overline{X}^*(l), ... \overline{X}^*(Q_k) \qquad \textbf{(3.3)}$$

Присоединим столбцы (3.2) к столбцам исходной матрицы ***X*** и для этой объединенной матрицы $\widetilde{X}$ *(N×(M+Q_k))* найдем решение QSAR-задачи МГУА-методом.

3.2.2 Составные фрагменты М-графов.

Определение. Одновременное присутствие в структурном спектре М-графа **G** двух неэквивалентных фрагментов ***f_1*** и ***f_2*** назовем ***2-составным фрагментом*** $F[2] = f_1 \otimes f_2$. Одновременное присутствие в структурном спектре М-графа **k** неэквивалентных фрагментов ***f_1,f_2,...,f_k*** назовем ***k-составным фрагментом*** $F[k] = f_1 \otimes f_2 \otimes ... \otimes f_k$.

Утверждение3.3 Пусть задан **k**-составной фрагмент ***f***= $F[k] = f_1 \otimes f_2 \otimes ... \otimes f_k$. Тогда вектор-столбец $\overline{X}[f]$, соответствующий числу повторений составного фрагмента ***f*** в обучающей выборке равен поэлементному произведению соответствующих столбцов матрицы ***X***:

$$\left(\overline{X}[f]\right)_i = \left(\overline{X}_1 \otimes \overline{X}_2 \otimes ... \otimes \overline{X}_k\right)_i = \prod_{j=1}^{k} \left(\overline{X}_j\right)_i \text{ , } \textbf{\textit{i=1,...,N}}$$

где $\left(\overline{a}\right)_i$ обозначает **i**-ю компоненту вектора $\overline{a}$.

Утверждение 3.4 Выберем в QSAR-матрице ***X*** набор столбцов с номерами ***(j_1,j_2,...,j_k)*** и поэлементно перемножим их: $\overline{X}_z = \overline{X}_{j_1} \otimes \overline{X}_{j_2} \otimes ... \otimes \overline{X}_{j_K}$. Тогда вектор-столбец $\overline{X}_z$ будет указывать число повторений **k**-составного фрагмента ***f*** в обучающей выборке, где ***f***= $F[k] = f_{j_1} \otimes f_{j_2} \otimes ... \otimes f_{j_k}$.

Справедливость утверждений 3.3 и 3.4 вытекает из построения матрицы ***X***, ***j***-й столбец которой указывает число повторений фрагмента ***f_j*** в обучающей выборке. Утверждение 3.4 позволяет построить МГУА-схему автоматического поска ***k***-составных фрагментов, адаптированных для некоторого свойства ***Y***.

Первая селекция построения “составных фрагментов”.

По матрице ***X*** построим ***M*(M-1)/2*** функций:

$$\overline{y} = F(X_{j_1}, X_{j_2}) + \overline{\varepsilon} = b_0 + b_1 \cdot (\overline{X}_{j_1} \otimes \overline{X}_{j_2}) + \overline{\varepsilon}\,,\ \Phi = (\overline{\varepsilon}, \overline{\varepsilon}) = \sum_{i=1}^{N} \varepsilon_i^2 \to \min\,,$$

которые упорядочим по возрастанию критерия ***Φ***. Отберем лучшие Q_1 уравнений с учетом критерия максимальной попарной коррелированности векторов с порогом $R_T(1)$. На основе отобранных уравнений сформируем новые столбцы признаки:

$$\overline{S}_l(1) = \overline{X}_{j_1}(l) \otimes \overline{X}_{j_2}(l)\,, \qquad l=1,...,Q_1$$

где $\overline{X}_{j_1}(l)$ и $\overline{X}_{j_2}(l)$ - вектор-столбцы, вошедшие в ***l***-ое отобранное уравнение.

***(k+1)**-ая селекция “составных фрагментов”.*

Пусть на ***k***-й селекции получено Q_k столбцов-признаков:

$$\{\overline{S}(k)_l\}\ \ l=1,...,Q_k;\ \ Corr(\overline{S}(k)_l, \overline{S}(k)_q) < R_T(k)\,,\ \ 1 \le l < q \le Q_k.$$

Построим $(Q_k * M)$ уравнений:

$$\overline{y} = \overline{F}(\overline{X}_j, S(k)_l) + \overline{\varepsilon} = b_0 + b_1 \cdot (\overline{X}_j \otimes S(k)_l) + \overline{\varepsilon}\,,\ \Phi = (\overline{\varepsilon}, \overline{\varepsilon}) \to \min$$

Упорядочим уравнения по возрастанию критерия ***Φ*** и отберем Q_{k+1} первых уравнений с учетом критерия максимальной попарной коррелированности с порогом $R_T(k+1)$. На основе отобранных уравнений сформулируем столбцы-признаки:

$$\overline{S}(k+1)_l = \overline{X}_{j_1}(l) \otimes \overline{S}(k)_{q_1}\,, \qquad l=1,...,Q_{k+1},$$

где $\overline{X}_{j_1}(l)$ и $\overline{S}(k)_{q_1}$ - вектор-столбцы, вошедшие в *l*-е отобранное уравнение.

Критерием остановки является проведение заданного числа селекций ***K***. “Раскроем скобки” для столбцов-признаков $\{S(k)_l\}$, $l=1,...,Q_k$, полученных на ***k***-ой селекции. Для фиксированного ***l*** имеем:

$$\overline{X}^*(l) = \overline{S}(k)_q = \overline{X}_{j_1}(l) \otimes \overline{S}(k-1)_{q_1} = ... \qquad \textbf{(3.4)}$$

$$= \overline{X}_{j_1} \otimes \overline{X}_{j_2} \otimes \overline{S}(k-2)_{q_2} = ... = \overline{X}_{j_1} \otimes \overline{X}_{j_2} \otimes ... \otimes \overline{X}_{j_{k+1}}\,,\ \ l=1,...,Q_k$$

Согласно утверждению 3.4, произведение (3.3) соответствует составному фрагменту $p = \{f_{j_1}, f_{j_2}, ..., f_{j_{k+1}}\}$. МГУА-построение “составных фрагментов” обеспечивает их адекватность (для заданного исходного списка фрагментов)

при аппроксимации свойства **Y**. Для каждого ***l (l=1,...,Q_k)*** будут получены отдельные “составные фрагменты”, которым будут соответствовать вектора-столбцы: $\overline{X}^*(1), \overline{X}^*(2),..., \overline{X}^*(l),... \overline{X}^*(Q_k)$. Присоединим эти столбцы к столбцам исходной матрицы ***X*** и и для этой объединенной матрицы $\widetilde{X}$ *(N×(M+Q_k))* найдем решение QSAR-задачи МГУА-методом.

3.2.3 Относительные фрагменты М-графа.

Определение. Вектор-столбец $\overline{X}[p]$, соответствующий числу повторений “относительного фрагмента ***p***” (фрагмент ***f_1*** относительно фрагмента ***f_2***) в обучающей выборке, равен отношению элементов соответствующих столбцов матрицы ***X***. Для того, чтобы охватить случаи, когда фрагмент ***f_2*** отсутствует в графе, запишем это определение так:

$$\left(\overline{X}[p]\right)_i = \left(\overline{X}_1 \div \overline{X}_2\right)_i = \frac{\left(\overline{X}_1\right)_i}{\left(\overline{X}_2\right)_i + 1}, \quad \textbf{\textit{i=1,...,N}};\ \ \overline{X}_1 = \overline{X}[f_1];\ \overline{X}_2 = \overline{X}[f_2];$$

где $\left(\overline{a}\right)_i$ обозначает ***i***-ю компоненту вектора $\overline{a}$.

Утверждение 3.5 Выберем в матрице *X* два столбца с номерами ***j_1,*** и ***j_2***, и вычислим их отношение: $\overline{X}_z = \overline{X}_{j_1} \div \overline{X}_{j_2}$. Тогда вектор-столбец $\overline{X}_z$ будет указывать число повторений “относительного” фрагмента *p* в обучающей выборке.

Справедливость утверждения 3.5 вытекает из построения матрицы ***X, j***-й столбец которой указывает число повторений фрагмента ***f_j*** в обучающей выборке. 8-й позволяет построить схему автоматического поиска “относительных” фрагментов, адаптированных для некоторого свойства. Поиск “относительных” фрагментов будем проводить только на одной (первой) селекции.

Селекция поиска “относительных фрагментов”.

По матрице ***X*** построим ***M*(M-1)/2*** функций

$$\overline{y} = F(X_{j1}, X_{j2}) + \overline{\varepsilon} = b_0 + b_1 \cdot \left(\overline{X}_{j_1} \div \overline{X}_{j_2}\right) + \overline{\varepsilon},\ \Phi = (\overline{\varepsilon}, \overline{\varepsilon}) = \sum_{i=1}^{N} \varepsilon_i^2 \to \min,$$

которые упорядочим по возрастанию критерия ***Φ***. Отберем лучшие ***Q_1*** уравнений с учетом критерия максимальной попарной коррелированности векторов с порогом ***$R_T(1)$***. На основе отобранных уравнений сформируем новые столбцы признаки:

$$\overline{S}_l(1) = \overline{X}_{j_1}(l) \div \overline{X}_{j_2}(l), \quad l=1,\ldots,Q_1 \qquad (3.5)$$

где $\overline{X}_{j_1}(l)$ и $\overline{X}_{j_2}(l)$ - вектор-столбцы, вошедшие в ***l***-е отобранное уравнение. Присоединим столбцы (3.4) к столбцам исходной матрицы ***X***:

$$\widetilde{X} = X \oplus \overline{S}_1(1) \oplus \ldots \oplus \overline{S}_{Q_1}(1)$$

и для этой объединенной матрицы $\widetilde{X}$ *(N×(M+Q1))* найдем решение QSAR-задачи МГУА-методом.

Рассмотренный способ поиска описания М-графа в виде структурных фрагментов разных типов позволяет автоматически генерировать наборы дескрипторов для *любых классов* химических веществ. Полученные дескрипторы являются *универсальными* и легко *интерпретируемыми*. Данный способ (за счет варьирования правил маркировки атомов в цепочках и выбора соответствующего описания обучающей выборки М-графов) позволяет добиваться получения набора дескрипторов *наиболее адекватно описывающих исследуемое свойство*. Представление результатов поиска в виде канонически закодированных фрагментов и наборов правил для генерации кодов позволяет хранить результаты поиска в виде *QSAR базы знаний*, внутренняя организация которой рассматривается ниже.

3.3 Генерация молекулярных графов с использованием эволюционных алгоритмов

3.3.1 Переборное решение обратной QSAR-задачи

Построение компьютерных моделей “структура-свойство” в рамках тех или иных математических подходов и методов [319,320,321] предполагает нахождение химических структур, которые можно назвать “активными в рамках построенной модели прогноза”. Это означает, что модель прогноза, получив на вход такую структуру, выдаст высокое значение ее активности. Если построенная модель прогноза является “простой” с математической точки зрения, например, однопараметрической, то возможно полное аналитическое решение “обратной задачи” нахождения всех химических структур, удовлетворяющих построенной модели прогноза [322]. Если же математическая модель прогноза представляет собой многопараметрическую функциональную зависимость, которая является нелинейной или представляет собой объединение нескольких функциональных зависимостей

(моделей) [323], то практическое решение “обратной задачи” может быть сведено к скринингу (просеиванию через математическую модель) химических соединений из некоторых классов, интересных для химика[324].

Сложность задачи состоит в том, что последовательность соединений для скрининга должна удовлетворять двум противоречивым условиям. С одной стороны, ограниченные вычислительные ресурсы заставляют сужать множество подаваемых на вход модели структур. В то же время, желательно получать соединения не только уже известных классов, но и принципиально новые. Таким образом, актуальной является разработка таких методов генерации, которые порождают “потенциально активные” структуры, с большой вероятностью обладающие заданным свойством.

Один из способов генерации молекул заключается в комбинаторном переборе всех структур по заданным брутто-формулам [325]. Другой способ предполагает более активное участие пользователя, который задает “основу” - базовый фрагмент молекулы, списки заместителей и указывает места присоединения этих заместителей к основе[326]. Существенный недостаток этого метода заключается в том, что класс получаемых соединений жестко определяется выбором “основы”.

3.3.2 Схема эволюционной генерации молекулярных графов

Предлагается новый метод генерации, позволяющий обойти эту проблему. Допустим, на молекулярно-химических графах введена метрика *M*, основное назначение которой — ограничить множество МХ-графов, поступающих на вход модели прогноза. Тогда можно использовать следующий алгоритм построения искомой последовательности графов. Выберем из обучающего множества структур два молекулярно-химических графа G_S и G_T, обладающих заданными значениями интересующего нас свойства. Будем строить последовательность графов $\mathbf{G_i}$, такую, что

$$\mathbf{G_1 = G_S,\ d(G_T, G_n) = 0;\ d(G_{i+1}, G_T) < d(G_i, G_T),\quad i = \overline{1,n};} \qquad (3.5)$$

где $\mathbf{d(G_1,G_2)}$ - расстояние между графами $\mathbf{G_1,G_2}$ в выбранной метрике.

Последовательность $\mathbf{\{G_i\}}$ можно получить с использованием эволюционного алгоритма[327],который неформально имеет следующую схему. Введем на графах следующие элементарные операции: добавить/убрать атом/ребро, переименовать атом, изменить кратность связи, переставить местами

заместители. Пусть на 1-ом шаге сформировано **K** графов-претендентов **$\{G_i^k\}$.** Обработку каждого **k**-го графа ($k = \overline{1,K}$) назовем *ветвью*. Применим к графу G_i^k одну из возможных элементарных операций. Получим граф-претендент G'_{i+1} следующего шага. При выполнении условия **(3.5)** добавим его в список $LG_k(i)$ претендентов **k**-й ветви. Формирование списка LG_k будем проводить до исчерпания возможности применения элементарных операций к различным структурным элементам (атомам, ребрам, заместителям) графа G_{ik}. После обработки всех ветвей сформируем общий список претендентов:

$$LG(i) = \bigcup_{k=1}^{K} LG_k(i) \qquad (3.6)$$

Из **LG(i)** отберем **K** "лучших" (в смысле близости к G_B) графов G_{i+1}^k, каждый из которых породит "ветвь" **(i+1)** шага (селекции). Для построения графов **(i+2)-**ой селекции повторим эту процедуру для каждого графа G_{i+1}**.** Условие **(3.5)** гарантирует "сходимость" последовательности к G_B. Таким образом, задача генерации МХ-графов сводится к построению метрик (в рамках данной QSAR-модели), адекватных исследуемому свойству (биологической активности).

3.3.3 Алгоритм эволюционной генерации

Выберем из обучающей СБД две активные структуры: G_S - стартовый граф и G_T - целевой граф. С помощью эволюционного алгоритма [327] построим список структур G_i, , таких что:

$$G_1 = G_S,\ d(G_T, G_n) = 0;\ d(G_{i+1}, G_T) < d(G_i, G_T),\quad i = \overline{1,n};$$

где $d(G_1,G_2)$ - расстояние между графами G_1,G_2 в выбранной метрике. Последовательность $\{G_i\}$ является результатом работы алгоритма.

Начальное семейство решений LG_1 состоит из одного стартового графа G_S. Введем на молекулярных графах следующие элементарные операции преобразования $\{ T_i \}$: добавить или удалить ребро, добавить или удалить атом, переименовать атом, изменить тип связи. Опишем эти операции полагая, что **V(G), E(G)** - соответственно множество вершин и ребер графа **G**; **p(v)** - степень вершины **v**; вставляемые вершины имеют метку атома углерода, а вставляемые ребра - метку одинарной связи.

***T_{DE}: Операция удаление ребра**: $G_2=T_{DE}(G_1)$;*

$V(G_2) = V(G_1)$; $E(G_2) = E(G_1) \setminus \{ e \}$, $e \in E(G_1)$

Ограничение: связность графа не должна нарушаться, т.е. могут удаляться только "кольцевые" ребра;

Вероятность применения P_{DE} :

$P_{DE} = 0$, если $| E(G_T) | > | E(G_1) |$;

$P_{DE} = (|E(G_1)| - | E(G_T) |) / | E(G_T) |$, если $| E(G_T) | \leq | E(G_1) |$,

где $| E(G) |$ – число ребер в графе G.

***T_{DV1}: Операция удаление вершины** степени 1 и его ребра: $G_2=T_{DV1}(G_1)$;*

$V(G_2) = V(G_1) \setminus \{ v \}$, $v \in V(G_1)$; $p(v)=1$

$E(G_2) = E(G_1) \setminus \{ e \}$, $e \in E(G_1)$, и e инцидентно v.

Ограничение: $p(v)=1$;

Вероятность применения P_{DV1} :

$P_{DV1} = 0$, если $| V(G_T) | > | V(G_1) |$;

$P_{DV1} = (|V(G_1)| - | V(G_T) |) / | V(G_T) |$, если $| V(G_T) | \leq | V(G_1) |$,

где $| V(G) |$ есть число вершин в графе G.

***T_{DV2}: Операция удаление вершины** степени 2 и одного ребра: $G_2=T_{DV1}(G_1)$;*

$V(G_2) = V(G_1) \setminus \{ v \}$, $p(v)=2$;

$E(G_2) = (E(G_1) + \{ (u,w) \}) \setminus \{ (v,u), (v,w) \}$,

где вершины v, u ,w $\in V(G_1)$; ребра (v,u), (v,w) $\in E(G_1)$,.ребро (u,w) $\notin E(G_1)$

Ограничение: ребра (v,u), (v,w) имеют метки одинарной связи; $p(v)=2$.

Вероятность применения P_{DV2} : $P_{DV2} = P_{DV1}$

***T_{IE}: Операция вставки ребра,** инцидентного вершине v: $G_2=T_{IE}(G_1)$;*

$V(G_2) = V(G_1)$; $E(G_2) = E(G_1) + \{ (v,u) \}$,

где вершины v,u $\in V(G_1)$, ребро (v,u) $\notin E(G_1)$;

Рассматривается вершина v, к которой присоединяется вершина u, выбираемая случайно из "некольцевых" вершин графа G_1.

Ограничение: степени вершин p(v), p(u) в графе G_2 не должны превышать валентности атомов v, u;

Вероятность применения P_{IE} :

P_{IE} =0, если | $E(G_T)$ | < | $E(G_1)$ |;

P_{IE} = (|$E(G_T)$ | - | $E(G_1)$ |) / | $E(G_T)$ |, если | $E(G_T)$ | $\geq$ | $E(G_1)$ |,

T_{IV1}: Операция вставки вершины *степени 1, инцидентной вершине v: $G_2=T_{IV1}(G_1)$;*

$V(G_2) = V(G_1) + \{ v \}$, $v \notin V(G_1)$; $p(v)=1$;

$E(G_2) = E(G_1) + \{ (v,u) \}$, $u \in V(G_1)$.

Рассматривается вершина v, к которой присоединяется новая вершина u.

Ограничение: степень вершины p(v) в графе G_2 не должна превышать валентности атомов v;

Вероятность применения P_{IV1} :

P_{IV1} =0, если | $V(G_T)$ | < | $V(G_1)$ |;

P_{IV1} = (|$V(G_T)$ | - | $V(G_1)$ |) / | $V(G_T)$ |, если | $V(G_T)$ | $\geq$ | $V(G_1)$ |,

T_{IV2}: Операция вставки вершины *степени 2 : $G_2=T_{IV2}(G_1)$;*

$V(G_2) = V(G_1) + \{ v \}$, $p(v)=2$;

$E(G_2) = (E(G_1) \setminus \{ (u,w) \}) + \{ (v,u), (v,w) \}$,

где вершины u ,w $\in V(G_1)$; вершины v $\notin V(G_1)$;

ребра (v,u), (v,w) $\notin E(G_1)$,.ребро (u,w) $\in E(G_1)$

Вероятность применения P_{IV2} : $P_{IV2} = P_{IV1}$

T_{ME}: Операция перемещения ребра*: $G_2=T_{ME}(G_1)$;*

$V(G_2) = V(G_1)$; $E(G_2) = (E(G_1) \setminus \{ (v,u) \}) + \{ (v,w) \}$;

ребра (v,u), (u,w) $\in E(G_1)$, ребро (v,w) $\notin E(G_1)$,

Ограничение: метка перемещаемого ребра (v,u) - одинарная связь,. p(u) $\geq$ 3

Вероятность применения P_{ME} :

P_{ME} =0, если | $E(G_T)$ | $\neq$ | $E(G_1)$ |;

P_{ME} =0, если p(u) = 2;

P_{ME} = (p(u) -2) / | $E(G_T)$ |, если | $E(G_T)$ | = | $E(G_1)$ | и p(u) > 2,

T_{LV}: Операция переименования вершины*: $G_2=T_{LV}(G_1)$;*

$V(G_2) = V(G_1)$, $E(G_2) = E(G_1)$, VLabel(v,G_1)=A1, VLabel(v,G_2)=A2, A1 $\neq$ A2;

где VLabel(v,G) - имя элемента вершины (атома) v в графе G;

Рассматривается вершина v с именем A1 (#VLabels(A1,G_1) ≠ #VLabels(A1,G_T)), для которой выбор нового имени A2 проводится случайно при условии #VLabels(A2,G_1) < #VLabels(A2,G_T)

Вероятность применения P_{LV} :

P_{LV} = 0, если #VLabels(A1,G_1) = #VLabels(A1,G_T)

P_{LV} = | #VLabels(A1,G_1) - #VLabels(A1,G_T) | / #VLabels(A1,G_T),

если #VLabels(A1,G_1) ≠ #VLabels(A1,G_T);

где #VLabels(A1,G) - число атомов в графе G с именем A1.

T_{LE}: Операция переименования ребра*: $G_2=T_{LE}(G_1)$;*

$V(G_2) = V(G_1)$, $E(G_2) = E(G_1)$,

ELabel(e,G_1)=E1, ELabel(e,G_2)=E2, E1 ≠ E2;

где ELabel(e,G) - метка ребра e в графе G;

Рассматривается ребро e с меткой E1 (#ELabels(E1,G_1) ≠ #ELabels(E1,G_T)), для которого выбор новой метки E2 проводится случайно при условии

#ELabels(E2,G_1) < #ELabels(E2,G_T)

Вероятность применения P_{LE} :

P_{LE} = 0, если #ELabels(E1,G_1) = #ELabels(E1,G_T)

P_{LE} = | #ELabels(E1,G_1) - #ELabels(E1,G_T) | / #ELabels(E1,G_T),

если #ELabels(E1,G_1) ≠ #ELabels(E1,G_T);

где #ELabels(E1,G) - число ребер в графе G с меткой E1.

Таким образом, вероятности применения операций преобразования графа родителя $\mathbf{G_1}$ *динамически изменяются* в зависимости от его размера (т.е. от числа его вершин и ребер) относитетельно целевого графа $\mathbf{G_T}$. Преобразование графа-родителя $\mathbf{G^j}$ (на шаге **j**) будем проводить следующим образом. Для каждой вершины **v** графа $\mathbf{G^j}$ применяются (с соответствующими вероятностями) операции: вставки вершины $\mathbf{T_{IV1}}$, $\mathbf{T_{IV2}}$, удаления вершины $\mathbf{T_{DV1}}$, $\mathbf{T_{DV2}}$, вставки ребра $\mathbf{T_{IE}}$ и переименования вершины $\mathbf{T_{LV}}$. Для каждого ребра e графа $\mathbf{G^j}$ применяются (с соответствующими вероятностями) операции: удаления ребра $\mathbf{T_{DE}}$, перемещения ребра $\mathbf{T_{ME}}$ и переименования ребра $\mathbf{T_{LE}}$. Если применение какой-либо операции произошло, то формируется граф-

потомок $\mathbf{G'^{j+1}}$ следующего шага. При выполнении условия $d\left(G'^{j+1}, G_T\right) < d\left(G^j, G_T\right)$ граф-потомок $\mathbf{G'^{j+1}}$ добавляется в список потомков $\mathbf{LG^{j+1}}$ графа $\mathbf{G^j}$, и продолжается просмотр и преобразование вершин и ребер графа-родителя $\mathbf{G^j}$.

Условие (3.5) гарантирует “сходимость” последовательности к целевому графу $\mathbf{G_T}$ (т. е. к графу, совпадающему с $\mathbf{G_T}$ в выбранной метрике). Построение такого графа $\mathbf{G_n}$ ($\mathbf{d(G_n,G_T) = 0}$) является условием остановки эволюционного алгоритма. Результатом генерации будем считать объединение всех списков графов-родителей LP^j: $LP = \bigcup_{i=1}^{n} LP^j$, где **n** - число шагов генерации.

Варианты оптимизации алгоритма[328]. Для эффективной реализации алгоритма целесообразно внести ряд ограничений, направленных на отказ от заведомо неприемлемых преобразований графа:

1. не допускаются модификации графа, которые приводит к нарушению валентности атомов;
2. при удалении ребер контролируется сохранение связности графа.

Повысить эффективность генерации можно, если определить структурное сходство стартового $\mathbf{G_S}$ и целевого $\mathbf{G_B}$ графов. Для этого найдем их наибольший общий фрагмент [329] $G'=MCS(G_S, G_B)$, т.е. такой молекулярный граф, что

1. $MCS(G_S\ G_B) \subseteq G_S$, ; $MCS(G_S\ G_B) \subseteq G_B$
2. $\forall G: (G \subseteq G_S, G \subseteq G_B) \Rightarrow (G \subseteq MCS(G_S\ G_B))$

Если наибольший общий фрагмент ***G’*** найден, то операции преобразования не должны затрагивать подграф ***G’*** — т. е. не должны модифицироваться атомы и ребра, которые уже обеспечивают структурное сходство графа-родителя и целевого графа.

Целесообразно предусмотреть следующий режим работы алгоритма генерации: вначале рассматриваются и преобразуются только “каркасные” графы (т. е. с графами без символьной разметки атомав и ребер). Тогда и фрагменты, относительно которых формируется структурный спектр графов, не будут содержать меток вершин и ребер. Это значительно позволит ускорить

вычисление расстояний между графами, а среди элементарных преобразований останутся только операции вставки, удаления и перемещения атомов и ребер. После генерации цепочки каркасных графов минимальной длины необходимо будет расставить (используя различные варианты расстановок) метки на атомах и связях сформированых графов.

Достоинства и недостатки различных вариантов “основного” алгоритма, а также оптимальные значения параметров могут быть определены в ходе экспериментального тестирования алгоритма на существующих моделях QSAR-прогноза. Очевидно, что наиболее важным параметром алгоритма генерации является выбор списка фрагментов ***H***, относительно которого формируется структурный спектр графов. Таким образом, наибольший эффект с точки зрения генерации графов, обладающих в заданной QSAR-модели "высокой" активностью, может дать оптимизация метрики относительно выбора ***H***. Кроме этого, второй уровень оптимизации метрики возможен за счет использования генетического алгоритма для “настройки” весовых коэффициентов w_j на основе информации об активности структур обучающего множества [330]. Программный модуль, реализующий описанный алгоритм, будет включен в QSAR-систему. В настоящее время проводится его реализация и тестирование совместно с аспирантом факультета Вычислительной математики и кибернетики МГУ А.А.Чернобаевым.

3.4 Выбор метрик М-графов с помощью генетического алгоритма.

Опишем методику селекции метрик молекулярных графов химических соединений на основе генетического алгоритма[331]. Один из способов построения метрики основан на нахождении максимального общего фрагмента графов[329]. Предлагается другой метод, базирующийся на разбиении графов обучающей выборки на фрагменты данного типа. Метрики вводятся на признаках в пространстве большой размерности, где признаком является число вхождений в данный молекулярный граф фрагментов заданной сложности. Рассмотрены два варианта критерия отбора метрик: первый - на основе информационной меры Шеннона, второй - на непараметрическом оценивании свойств веществ.

3.4.1 Схема отбора метрики

Пусть задана обучающая СБД. Зададим сложность фрагментов и маркировку атомов, и тем самым определим вид индуктивного структурного спектра, который будет использован в построении метрики. Сформируем матрицу "молекула-признак" $\mathbf{X=(X}_{ij}\mathbf{)}$, $i=1,...,N$; $j=1,...,M$; где $\mathbf{X}_{ij}$-число повторений j-го фрагмента в i-ой молекуле, **N**-число молекул, **M**-число дескрипторов (фрагментов).. Затем используется одна из следующих метрик:

1. "Прямоугольная" (манхэттенская) метрика

$$M_W(G_1, G_2) = \sum_{j=1}^{M} w_j \cdot \left| x_j^{(1)} - x_j^{(2)} \right|$$

2. Евклидова метрика

$$D_W^2(G_1, G_2) = \sum_{j=1}^{M} w_j \cdot \left(x_j^{(1)} - x_j^{(2)} \right)^2$$

3. Метрика Хэмминга

$$H(G_1, G_2) = \sum_{j=1}^{M} w_j \cdot h\left(x_j^{(1)}, x_j^{(2)} \right)^2, \text{ где } h(a,b) = \begin{cases} 1, & a > 0,\ b > 0 \\ 1, & a = 0,\ b = 0 \\ 0 & \end{cases}$$

3.4.2 Генетический алгоритм

Будем проводить перебор различных вариантов метрики с помощью генетического алгоритма [327]. Для каждого варианта метрики проводится кластеризация молекул СБД (например, с помощью алгоритма k-средних). Качество полученного разбиения молекул обучающей СБД (кластерная структура) оценивается с точки зрения ее адекватности классификации молекул по данному свойству. На каждом селекции отбираются лучшие метрики (т.е. метрики, порождающие лучшие разбиения), которые и будут участвовать в генерации новых вариантов.

"Ген", определяющий текущее решение задачи, представляет собой вектор $\mathbf{w} \in \mathbf{R}^M$, где w_j - весовой коэффициент j-го дескриптора в метрике (например, евклидовой). На начальном этапе формируется случайное семейство метрик (решений), состоящее из **Q** генов. Для каждого члена семейства (родителя) формируются новые метрики (потомки). Из полученных

потомков отбираются Q “лучших” (согласно критерию отбора) - это будет семейство решений-родителей для следующего шага оптимизации.

C*пособы формирования потомков.*

(1) “Мутация”: проводится изменение параметров родителя в одном компоненте (в нашем случае - это изменение одного из весовых коэффициентов) : $w_j(k+1) = w_j(k) + \Delta w$, где ***k*** - номер текущего шага. Частота мутаций и их величина могут выбираться различными способами.

(2) “Кроссовер”: для двух данных родителей формируются два потомка, содержащие часть гена одного родителя, а часть - другого; например, следующим образом. Сформируем случайный двоичный вектор $\bar{b} = (b_1, ..., b_M)$; $b_j \in \{0,1\}$; $j=1,...,M$. Тогда пусть первый потомок получит ***j***-ый вес первого родителя, если $b_j = 1$, и вес второго родителя, если $b_j = 0$; а второй - наоборот. Выбор родителей также может быть различным:

(а) все **Q** родителей разбиваются на пары, каждая пара порождает двух потомков, получаем **Q** потомков всего множества;

(б) рассматриваются все возможные родительские пары (**Q*(Q-1)/2** пар), которые порождают **Q*(Q-1)** потомков.

*Критерий отбора***.** Отбор может проводится: (а) только среди потомков; (б) среди потомков и родителей. Второй вариант предпочтителен, т.к. найденное оптимальное решение-родитель будет оставаться во всех селекциях. Рассмотрим два варианта критерия отбора.

Информационная мера Шеннона. Пусть заданное свойство разбивает все множество молекул на "активнве" и "неактивные", и проведена кластеризация объектов. Пусть имеем *k* кластеров, каждый из которых состоит из g_i графов **(*i*=1...*k*)**, где $\sum_{i=1}^{k} g_i = N$. Пусть в ***i***-ом кластере имеется a_i “активных” и b_i “неактивных” молекул, $a_i + b_i = g_i$.Тогда качество *i*-го кластера определим функцией

$\Phi_i = -p_i * \log(p_i) - q_i * \log(q_i)$,

где $p_i = a_i / g_i$; $q_i = b_i / g_i$ Критерий качества всего разбиения определим как $\Phi(k) = \sum_{i=1}^{k} W_i * \Phi_i$,где W_i - весовой коэффициент: $W_i = g_i / N$.

Именно эту функцию $\Phi(k)$ и выберем в качестве *критерия отбора генетического алгоритма*. Видно, что оценивающая функция $\Phi(k)$ достигает максимального значения, равного единице, в том и только в том случае, когда в каждом кластере присутствует ровно половина объектов первой группы и половина - второй ($p_i = q_i = 0.5$). Минимальное значение функции $\Phi(k)$, равное нулю, достигается, если в каждом кластере находятся объекты одной группы (кластеризация оказалась полностью успешной). Эту функцию будем минимизировать, подбирая весовые коэффициенты w_j. Пусть $w \in R^M$ вектор есть “ген”, определяющий решение задачи. Будем проводить на каждом шаге мутацию, т. е. изменение одного из коэффициентов по формуле $w_j(q + 1) = w_j(q) + \Delta w$, где q — номер шага, и $0 < \Delta w < 1$. Минимизация критерия в рамках генетического алгоритма приводит к построению метрик, в которых молекулы одного класса расположены компактно.

Непараметрическое оценивание свойств [332]. Определим “близость” двух молекулярных графов G_1 и G_2, как $M(G_1, G_2) = \frac{2 \cdot Common(G_1, G_2)}{S(G_1) + S(G_2)}$, где $Common(G_1, G_2)$ - число общих фрагментов графов G_1 и G_2, $S(G)$ - общее число фрагментов графа G. Пусть задано СБД, состоящая из N молекулярных графов $\{G_1, G_2, \ldots, G_N\}$ с известными значениями изучаемого свойства $v(G_i)$. Для оценки свойства нового графа G_{new} в списке заданных М-графов определим K максимально близких графов $\{G_{j1}, G_{j2}, G_{j3} \ldots G_{jK}\}$ к графу G_{new}. Оценку свойства v для графа G_{new} рассчитаем по следующей формуле: $v(G_{new}) = \sum_{k=1}^{K} W_k \cdot v(G_{jk})$, где $W_k = \frac{M(G_{new}, G_{jk})}{\sum_{k=1}^{K} M(G_{new}, G_{jk})}$.

Для построения функции оценки генетического алгоритма в качестве G_{new} рассмотрим последовательно все графы обучающей СБДи сформируем

вектор оценки свойств $\mathbf{v^*=(v^*(G_1), v^*(G_2),...,v^*(G_N))}$. Критерием качества метрики будем считать множественный коэффициент корреляции **R** заданного вектора свойств **v** относительно вектора $\mathbf{v^*}$ оценки свойств графов обучающего множества:

$$R^2 = \left(1 - \frac{\sum_{i=1}^{N} \left(v_i - v_i^* \right)^2}{\sum_{i=1}^{N} \left(v_i - \bar{v} \right)^2} \right)$$

Особенность предложенного подхода к генерации молекулярных графов заключается в возможности формирования соединений из новых химических классов. При этом из-за сохранения структурного сходства генерируемых графов с исходными повышается вероятность того, что модель прогноза будет к ним применима, и обратную задачу удастся решить. Программные модули, реализующие описанные варианты генетического алгоритма, будут включены в QSAR-систему. В настоящее время проводится их реализация и тестирование совместно с аспиранткой факультета Вычислительной математики и кибернетики МГУ Т.А.Юрченко и аспирантом Механико-математического факультета МГУ Д.Г.Вороной.

3.5 Выводы

Рассмотрена задача поиска наборов обобщенных фрагментов молекул для формирования количественных корреляций “структура-свойство” (QSAR) на основе самоорганизации линейных моделей по методу МГУА - метод группового учета аргументов. МГУА-подход хорошо зарекомендовал себя в практике построения QSAR-моделей при обработке индуктивных структурных спектров обучающих СБД, содержащих 100-500 молекулярных структур. Рассмотрены МГУА-схемы для построения и отбора составных несвязных фрагментов, формируемых на основе комбинаторного перебора базовых связных фрагментов заданной сложности. Описана схема эволюционного решения “обратной” QSAR-задачи путем генерации молекулярных графов, образующих путь минимальной длины между двумя целевыми структурами в пространстве помеченных графов. Преложенано схема селекции метрик молекулярных графов с использованием генетического алгоритма, основанная на разбиении графов обучающей СБД на фрагменты данного типа. Метрики вводятся на признаковых пространстве большой размерности, где признаком является число вхождений в М-граф заданных фрагментов. Двоичный вектор гена определяет, какие фрагменты включены в расчет метрики, а какие нет. Рассмотрены два варианта критерия отбора метрик: первый - на основе информационной меры Шеннона, второй - на непараметрическом оценивании свойств веществ.

Глава 4. Программная система для конструирования QSAR-моделей.

4.1 Основные типы данных и программные модули QSAR-системы.

В основу построения QSAR-системы были положены три принципа:

1. Система должна строиться на базе небольших программных модулей, которые могут выполняться как независимые ДОС-программы и обмениваться данными через внешние файлы. Это обеспечивает простую расширяемость системы и возможность независимой разработки модулей.

2. В системе вводятся стандарты на основные типы данных на диске, путем определения так называемых "коммуникационных форматов"

3. Программные модули строятся на основе объектов.

Основные типы данных, используемые при конструировании системы, представлены на рисунке 4.1. К ним относятся:

- "молекулярная структура",
- "обучающая выборка (база данных)",
- "матрица информативных признаков",
- "QSAR-уравнение"
- "QSAR-модель".

С этими данными работают основные программные модули:

- графический редактор структур
- редактор обучающей базы данных
- "дескрипторные модули"
- модуль, объединяющий признаки в матрицу описания
- модуль первичного отбора информативных признаков ("первичной селекции")
- модуль построения QSAR-уравнений
- модуль проведения "скользящего контроля"
- сервисные модули, включая интерфейсный блок.

Приведем краткую характеристику каждого модуля:

1). Графический редактор химических структур.

Назначение: в диалоговом режиме проводить формирование и редактирование структурных химических соединений путем рисования атомов и связей на экране дисплея.

Выход: таблица связи атомов в коммуникационном формате.

2). Редактор структурной обучающей базы данных.

Назначение: сопоставить и отредактировать числовую информацию о свойствах химического соединения, включая такие вычисляемые по молекуле характеристики, как молекулярный вес и брутто-формула.

Выход: файл в коммуникативном формате, задающий список свойств для каждой структуры обучающей выборки.

3). Блоки вычисления признаков.

Назначение: вычисление заданного класса инвариантов по молекулярной структуре, т.е. проведение "векторизации" структуры.

Вход: Обучающая структурная база данных, представленная в виде списка таблиц связи атомов.

Выход: для каждой молекулы файл со списком признаков, представленный в виде пары: (имя признака, значение).

Замечание: предполагается, что пользователи системы могут добавлять новые "дескрипторные блоки", спроектированные специально для расчетов данного свойства. Единственное ограничение на такие программные модули состоит в необходимости представления дескрипторов молекул в коммуникативном формате.

4). Блок формирования матрицы описания.

Назначение: собрать описания молекул, созданные различными дескрипторными блоками , в единую матрицу, каждый столбец которой буде соответствовать одному признаку.

Вход: список файлов, содержащих списки признаков для данной обучающей структурной базы данных.

Выход: матрица описания Y и список имен признаков, вошедших в матрицу описания.

5). Блок первичной селекции признаков.

Назначение: удаление из матрицы описания признаков-констант и неинформативных признаков с одновременным формированием новых (составных) признаков, удовлетворяющих критерию селекции.

Вход: критерий селекции признаков, матрица Y описания обучающей выборки.

Выход: матрица X информативных признаков-столбцов.

6). Блок построения линейного QSAR-уравнения.

Назначение: построение QSAR-уравнений по формуле (1), удовлетворяющих критерию селекции уравнений.

Вход: критерий селекции уравнений, матрица X информативных признаков.

Выход: список QSAR-уравнений (QSAR-модель), построенных на различных наборах-признаках X, и удовлетворяющих критерию селекции.

7). Блок "исследовательского анализа" QSAR-моделей.

Назначение: удаление/включение переменных в QSAR-уравнение с одновременным просчетом весовых коэффициентов. Проведение "скользящего контроля" QSAR-уравнений.

Вход: список QSAR-уравнений (QSAR-модель) и матрица описания Y.

Выход: список новых QSAR-уравнений, прошедших "ручную" настройку и проверку на прогностичность.

8). Блоки визуализации и сервисные блоки.

Вход: файлы с основными типами данных, в соответствующих коммуникативных форматах.

Выход: визуализация данных на дисплее компьютера или формирование и распечатка текстового файла с отчетом-протоколом.

Назначение: сервисные блоки позволяют пользователю в диалоговом режиме проводить следующие действия:

- просматривать обучающую структурную базу данных с одновременным маскированием (выключением из обучающей выборки) некоторых структур
- просматривать гистограммы распределения выбранного свойства
- просматривать дескрипторное описание структур
- просматривать матрицу описания Y и списки признаков
- просматривать матрицу X информативных признаков
- просматривать в графическом виде зависимость "наблюдаемые-предсказанные значения" свойства (с интерактивной визуализацией структурных формул молекул, соответствующих точкам на графике)
- просматривать вид QSAR-уравнений и формировать отчет-протокол

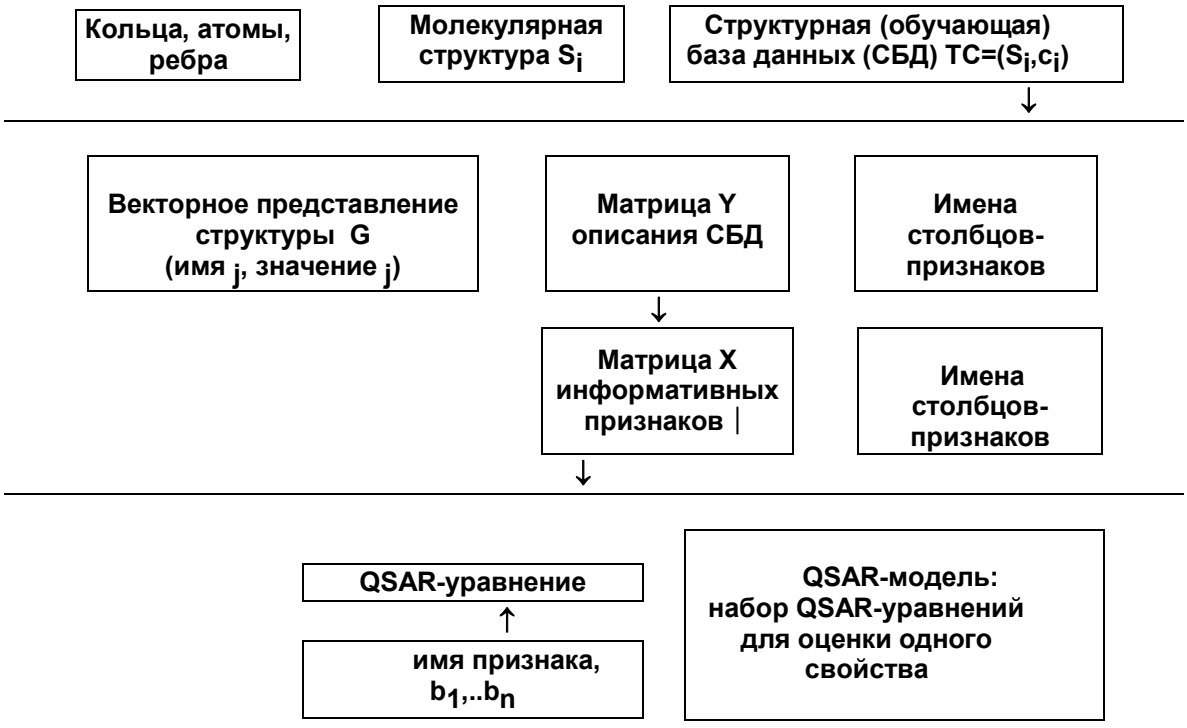


Рисунок 4.1 Основные типы данных системы прогнозирования свойств химических соединений.

4.2 Объектно-ориентированый подход к представлению структур данных в QSAR-системе.

4.2.1 Компьютерное представление М-графа.

Для создания приложений, работающих с химическими структурами, необходимо было иметь некий базовый инструментарий, позволяющий реализовывать алгоритмы работы с М-графами, описанные в главе 2, а именно:

- вводить М-графы, используя различные коммуникативные форматы
- распознавать основные структурные особенности М-графов (связности, циклы, ациклы)
- осуществлять поиск изоморфизма как М-графов, так и подграфов (с учетом химической специфики)
- строить на основе имеющихся программных объектов специфические алгоритмы обработки М-графов

Такой инструментарий был создан в виде программных библиотек на объектно-ориентированном расширении языка Borland Pascal. Выбор объектно-ориентированного представления в нашем случае неоднократно оправдал себя при изменении способов внутреннего представления М-графов, при создании новых классов алгоритмов, при организации взаимодействия между модулями. Унифицированность представления программных текстов, присущая объектно-ориентированному стилю, облегчает ориентацию среди значительного числа программных модулей, общий объем которых превышает 10,000 строк. В настоящее время широко распространены два основных метода представления М-графа:

- в виде линейного символьного кода;
- в виде какой-то модификации матрицы смежности М-графа.

Символьное представление М-графа наряду с положительными качествами (по-сути снимается "специфичность" графового представления, вся работа осуществляется с текстовыми строками) имеет существенные недостатки:

- кодирование с трудом поддается автоматизации и практически всегда осуществляется в ручную;

- зафиксированная система кодирования не является универсальной, появление нового класса соединений может приводить к существенной переработке правил кодирования;
- по символьному коду легко осуществить поиск изоморфизма целиком М-графа, но практически невозможно осуществить поиск М-графов, включающих заданные структурные фрагменты.

Все эти недостатки присущи, например, самой распространенной в настоящее время канонической системе кодирования органических соединений - системе Висвессера. Поэтому для построения универсальной системы алгоритмической обработки молекулярных структур выбор представления М-графа в виде матрицы связности является необходимым. Обычная матрица смежности является крайне неэффективной по размеру занимаемой памяти (NxN) и по скорости доступа к информации (для определения соседей атома i требуется просмотреть всю i-ю строку матрицы смежности). Поэтому для внутреннего представления М-графа был выбран список. М-граф представляется в виде списка пронумерованных объектов - атомов. Каждый атом содержит внутри себя всю информацию о пометках и о своих соседях.

На рисунке 4.2 приведены интерфейсные части (только данные) базовых объектов - Атома и Молекулы (М-графа).

```
const MaxNs = 6;  { Максимальное число соседей у атома }
type
   TNeighbour = record  { Информация о соседе }
            N,                  { Номер соседа в списке атомов М-графа }
            Conn: integer; { Размерность и тип связи }
            Aroma: boolean  { Является ли связь ароматической }
   end;

   TNeighbours = array[1..MaxNs] of TNeighbour;  { Список соседей атома }

   PAtom = ^TAtom;
   TAtom = object(TObject) { Объект - атом }
   Num,Charge,Coh,X,Y,Z: integer;  { Номер атома, заряд, номер связности, координаты для визуализации}
            nH: byte; { Число водородов, присоединенных к атому }
            Name: string[2]; { Название по периодической системе элементов}
            NB: TNeighbours; { Список соседей }
            Reserved: array [1..MaxNs] of integer;  { Зарезервированно для использования в различных приложениях }
            ..............................................
   end;

   TMolecule = object(TCollection) { Объект - молекула (М-граф) }
```

```
    Brutto: PString; { Строка - брутто-формула }
    Rings: PRings;  { Список колец М-графа }
    Systems: PRingSystems; { Список кольцевых систем }
    ..................................................
    end;
```

Рисунок 4.2 Интерфейсная часть базовых объектов - Атом и Молекула.

Поскольку библиотека реализовывалась с использованием стандартного модуля Objects, поставляемого в составе Borland Pascal, то объекты TAtom и TMolecule являются производными от объектов TObject и TCollection соответственно. Объект TObject является базовым объектом для всей иерархии объектов в Borland Pascal. Объект TCollection является неким "универсальным списком", хранящим ссылки на другие объекты и предоставляющим как последовательный (списочный), так и прямой (массивный) доступ к своим элементам. Иерархия базовых объектов приведена на рисунке 4.3.

Рисунок 4.3. Иерархия программных объектов: химический атом, М-граф, список свойств М-графа, М-граф со списком свойств. Жирным шрифтом выделены названия новых программных объектов.

В базовых объектах были реализованы основные алгоритмы работы с М-графами, рассмотренные выше в главе 2:

- распознавание структурных особенностей (связности. кольцевые системы, цепи, мостики, ароматические циклы)
- поиск изоморфизма графов и подграфов
- визуализация
- методы для управления модификацией структуры М-графов (вставка/удаление атомов, ребер)

На рисунках 4.4 - 4.11 приведен список методов в базовом объекте **TMolecule**. Имена процедур имеют мнемонические значения и указывают на функциональный смысл методов. В объяснениях используются общепринятые термины объектно-ориентированного программирования.

```
type
  TMolecule = object(TCollection)
  ..........................................................
   procedure FreeItem(Item: pointer); virtual;
   destructor Done; virtual;
   procedure PutItem(var S: TStream; Item: pointer); virtual;
   function GetItem(var S: TStream): pointer; virtual;
   procedure Store(var S: TStream);
   constructor Load(var S: TStream);
   constructor Init;
```

Рисунок 4.4 Интерфейсная часть (методы) базового объекта TMolecule. Переопределенные методы предка (TCollection**).**

Группа методов, приведенная на рисунке 4.4. обеспечивает стандартную обработку атомов М-графа, как элементов коллекции: удаление из памяти, запись/чтение атома из универсального объекта ввода/ вывода - потока; кроме того, реализует стандартные методы - конструктор для М-графа и полиморфный деструктор.

```
   procedure InitXXYY;
   procedure DoneXXYY;
   procedure ReScale; virtual;
   procedure Scale(X1,Y1,X2,Y2: integer; Grow: boolean);
   procedure Draw(X1,Y1,X2,Y2,Rad: integer);
   procedure ReDraw;
   procedure DrawFixed;
   procedure DrawScience(X1,Y1,X2,Y2,Rad: integer);
   procedure DrawScaleScience(X1,Y1,X2,Y2,Rad: integer);
   procedure ScienceDrawFixed;
   procedure ScienceReDraw; virtual;
   function Signature(Num: integer): string; virtual;
   procedure DrawBond(N1,N2,Conn: Integer; Aroma: boolean);
   procedure DrawAtom(Atom: PAtom); virtual;
   procedure CopyCoord(M: PMolecule);
   procedure SetAtomColors(N: Integer); virtual;
   procedure SetBondColors(N1,N2: Integer); virtual;
```

Рисунок 4.5 Интерфейсная часть (методы) базового объекта TMolecule**. Методы для визуализации М-графа.**

Группа методов, приведенная на рисунке 4.5 служит для изображения на дисплее М-графа в графическом виде (так, как принято изображать химические структуры). Координаты атомов М-графа соответствующим

образом масштабируются для изображения структуры в заданном прямоугольнике экрана. Виртуальные методы обеспечивают возможность для наследников TMolecule задания дополнительной символьной информации, изображаемой рядом с атомом. В настоящее время эта группа методов используется в основном в процессе отладки. В специальном наследнике реализованы методы для изображения структуры в окне MS Windows.

```
function DetectRings: boolean;
procedure DetectCoherences;
procedure DetectRibs;
procedure DetectSystems;
function BenzolRing(Num: integer): boolean;
function CarboRing(Num: integer): boolean;
procedure DetectAcs;
procedure DetectBenzol;
function Research: boolean; virtual;
function GetChain(N1,N2,MinLen: Integer): string;
```

Рисунок 4.6 Интерфейсная часть (методы) базового объекта TMolecule**. Методы для распознавания структурных особенностей М-графа.**

Группа методов, приведенная на рисунке 4.6. реализует алгоритмы распознавания таких структурных особенностей химической структуры, как наличие несвязанных фрагментов, колец, кольцевых систем, ациклов. Все структурные особенности представляются в виде объектов - упорядоченных списков номеров, входящих в них атомов.

```
procedure SetAromaRings;
procedure ClearAroma;
procedure SetAromaRing(R: PRing);
function AromaRing(R: PRing; var PiBonds: IntVector): boolean; virtual;
procedure DefinePi(var PiBonds: IntVector);
```

Рисунок 4.7. Интерфейсная часть (методы) базового объекта TMolecule**. Методы для распознавания ароматических систем**

Методы, приведенные на рисунке 4.7. предназначены для поиска ароматических кольцевых систем. Они основаны на алгоритме, описанном в второй главе.

```
function VinilAtom(Atom: PAtom): boolean; virtual;
function AromaAtom(Atom: PAtom): boolean; virtual;
function RingAtom(Atom: PAtom): boolean; virtual;
function AllRingAtom(Atom: PAtom): boolean; virtual;
function ChainAtom(Atom: PAtom): boolean; virtual;
function FusedAtom(Atom: PAtom): boolean; virtual;
function AromaFused(Atom: PAtom): boolean; virtual;
function AllAromaAtom(Atom: PAtom): boolean;virtual;
```

```
function Csp3Atom(Atom: PAtom): boolean; virtual;
function SubstAtom(Atom: PAtom): boolean; virtual;
function OnlyC(Atom: PAtom): boolean; virtual;
function PeptideAtom(Atom: PAtom): boolean; virtual;
```

Рисунок 4.8 Интерфейсная часть (методы) базового объекта TMolecule**. Методы для определения химических характеристик атомов в М-графе.**

Группа методов, приведенных на рисунке 4.8. позволяет определить некоторые "локальные" характеристики атомов, например, местоположение (в ацикле, кольце, в месте присоединения заместителя к кольцу, в пептидной цепочке и т.д.). Эти методы используют результаты работы методов поиска структурных особенностей всего М-графа.

```
function CalcBrutto: string;
function CalcMass: real;
function CalcAdvanceBrutto: string;
```

Рисунок 4.9 Интерфейсная часть (методы) базового объекта TMolecule**. Методы для вычисления простейших индексов М-графа: молекулярной массы и брутто-формулы.**

```
function AddAtom(Atom: TAtom): PAtom;
procedure DeleteAtom(N: Integer);
function CopyAtom(Atom: PAtom): PAtom;
procedure SetBond(N1,N2,Conn: integer);
function GetBond(N1,N2: Integer): Integer;
procedure SetAromaBond(N1,N2: Integer);
procedure DeleteBond(N1,N2: integer);
constructor MakeCopy(M: PMolecule);
procedure CopyM(M: PMolecule);
constructor Create(Atoms: integer);
```

Рисунок 4.10 Интерфейсная часть (методы) базового объекта TMolecule**. Методы управления структурой М-графа: добавления/удаления атомов/ связей, копирования М-графов.**

```
function Contain(SubM: PMolecule; var Fact: BoolVector):    boolean;
procedure Factorize(var M1,M2: PMolecule; Use: BoolVector);
constructor Extract(M: PMolecule; Use: IntVector; L: integer);
function MarkAtoms(Name: string; var Mark: IntVector; L: integer): integer;
function MarkCoherences(var Cohs: IntVector): integer;
procedure DeletePartByFact(var Fact: BoolVector; const NewName: NameType);
function DeletePart(Radical: PMolecule; NewName: NameType): boolean;
function Equal(M: PMolecule): boolean;
function AtomsEqual(SubM: PMolecule; var A1, A2: TAtom): boolean; virtual;
function BondsEqual(SubM: PMolecule; var A11, A12, A21, A22: TAtom): boolean;
virtual;
```

Рисунок 4.11 Интерфейсная часть (методы) базового объекта TMolecule**. Методы для определения изоморфизма графов, для поиска вхождений заданного подграфа (фрагмента) в М-граф.**

4.2.2 Компьютерное представление представление обучающей БД.

В обучающей выборке соединений находятся М-графы с приписанными им внешними свойствами (например, температура кипения соответствующего вещества, биологическая активность, спектр и т.д.). Необходимо уметь считывать М-графы с сопутствующей цифровой или текстовой информацией из файлов в различных форматах. Разнородная сопутствующая информация должна быть представлена в каком-то унифицированном виде.

Для представления обучающей выборки соединений был создан новый уровень иерархии объектов, представленный на рисунке 4.12

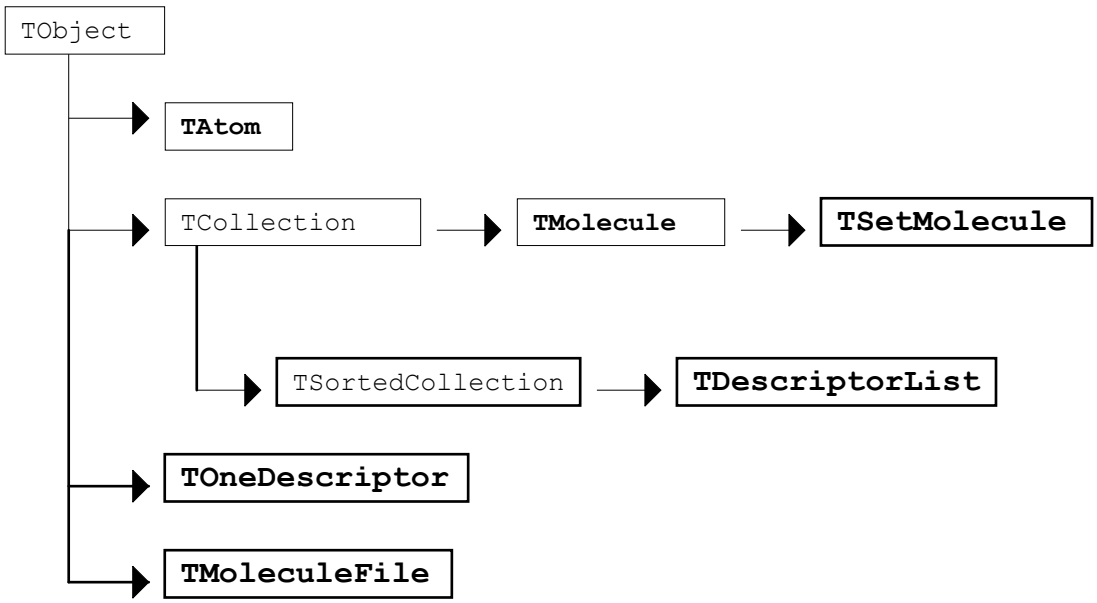


Рисунок 4.12 Иерархия объектов для представления "обучающей" БД.

Объекты **TOneDescriptor** и **TDescriptorList** используются для представления и доступа к сопутствующей информации. Для обеспечения универсальности сопутствующая информация представляется в виде текстовых строк, которые затем, исходя из контекста исполнения программы, подвергаются необходимым преобразованиям (в число, в набор чисел и т.д.). Данные содержатся в М-графе (объект **TSetMolecule**) как новое поле "**Descriptions**" и представляют собой список (**TDescriptorList**). Каждый элемент списка (**TOneDescriptor**) содержит название свойства и его значения

(поля **Name** и **Val** соответственно). Список свойств создан на основе стандартного объекта **TSortedCollection**, предоставляющего методы для быстрого (бинарного) поиска объектов, которые могут сравниваться по ключу. В качестве ключа в объекте **TDescriptorList** выступает имя свойства. На рисунке 4.13. представлены интерфейсные части этих объектов.

В ходе практической реализации были созданы методы-конструкторы для считывания М-графов с данными из нескольких текстовых форматов. Оказалось удобным объединить методы доступа к таким файлам в одном объекте - Молекулярном Файле (**TMoleculeFile**). Этот объект предоставляет унифицированные методы для определения числа структур в файле, для получения из файла структуры по ее номеру, для добавления структуры в файл и т.п. Вся обработка специальных форматов (даже многофайловых) осуществляется внутри этого объекта. Добавление внутрь этого объекта методов для работы с новым форматом автоматически делает возможным поддержку этого формата всеми приложениями, использующими **TMoleculeFile** для доступа к файлам структур. На основе объекта **TMoleculeFile** создан объект **TMoleculeDataBase** - Молекулярная База Данных, которая реализует СУБД химической ориентации с одновременной поддержкой (без конвертации) ранее использовавшихся текстовых форматов.

```
POneDescriptor = ^TOneDescriptor;
TOneDescriptor = object(TObject)
   Name,Val: Pstring;
    constructor Init(AName,AVal: string);
    constructor Load(var S: TStream);
    procedure Store(var S: TStream);
    destructor Done; virtual;
end;
PDescriptorList = ^TDescriptorList;
TDescriptorList = object(TSortedCollection)
    function KeyOf(Item: Pointer): Pointer; virtual;
    function Compare(Key1, Key2: Pointer): Integer; virtual;
    procedure PutItem(var S: TStream; Item: pointer); virtual;
    function GetItem(var S: TStream): pointer; virtual;
    function GetDes(N: Integer): OneDesPtr;
    function GetName(N: Integer): string;
    function GetVal(N: Integer): string;
    procedure SetName(N: integer; S: string);
    procedure SetVal(N: integer; S: string);
    constructor ReadDes(var f: text);
    constructor MakeCopy(D: PDescriptorList);
    function WriteDes(var f: text): boolean;
    function SearchDes(const Name: string): integer;
```

```
    function SearchDesByVal(Val: string): integer;
    constructor MergeDes(A1,A2: PDescriptorList);
    function AddDes(Name, Val: string): boolean;
    procedure DelDes(const Name: string);
  end;
  PSetMolecule = ^TSetMolecule;
  TSetMolecule=object(TMolecule)
    Descriptions: PDescriptorList;
    constructor ReadCpssText(var CBFile: Text);
    destructor Done; virtual;
    procedure WriteCPSS(var f: text);
    constructor Load(var S: TStream);
    procedure Store(var S: TStream);
    constructor MakeCopy(M: PMolecule);
    function GetDes(const Name: string): string;
    procedure PutDes(const Name, Val: string);
  end;
```

Рисунок 4.13 Интерфейсные части объектов, представляющих М-граф из обучающей выборки.

4.2.3 Компьютерное представление МГУА-данных.

Основные программные объекты, в рамках которых были реализованы схемы МГУА вычислений, представлены на рисунке 4.14.

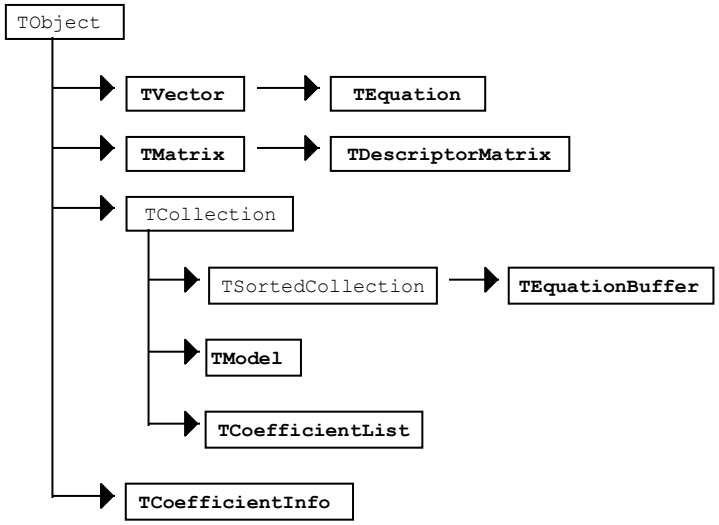


Рисунок 4.14 Иерархия программных объектов, реализующих МГУА-схему построения QSAR-моделей.

TEquation - программный объект, представляющий QSAR-уравнение. Он построен на основе объекта **TVector** и хранит вектор оценки свойств для структур обучающей выборки. Кроме того, он содержит список дескрипторов, входящих в уравнение, с их весовыми коэффициентами (**TCoefficientList**). По каждому дескриптору содерится информация, необходимая для его вычисления (**TCoefficietnInfo**): правила кодирования атомов, простой или сложный (обобщенный, составной или относительный) (в этом случае

включается информация о списке простых дескрипторов, входящих в сложный), его весовой коэффициент в уравнении. **TEquationBuffer** - объект, осуществляющий “отбор” QSAR-уравнений в ходе их построения на каждой селекции МГУА-процесса. **TModel** - объект, содержащий информацию по QSAR-модели, и позволяющий провести оценку свойства для нового М-графа.

4.3 Программная система BIBIGON.

Программная система **BIBIGON MATCH (B**asic **I**nstrument for **B**uilding **I**nteractive **G**eneration of **O**ptimized **N**etworks of **M**arked **AT**oms **Ch**ains) - предназначена для реализация “экспертного” QSAR-этапа. Система разработана в соавторстве с заведующим лабораторией Математической химии ИОХ РАН академиком Зефировым Н.С., заместителем заведующего этой же лабораторией к.х.н. Смоленским Е.А. и аспирантами Пономаревой Л.А. и Митюшевым Д.Ф.

Под управлением пользователя-эксперта система BIBIGON:

- проводит описание М-графа множеством структурных спектров (в виде маркированных цепочек атомов);
- вычисляет набор топологических, информационных, структурных индексов для использования в качестве “дополнительных” дескрипторов;
- строит QSAR-модели методом группового учета аргументов в автоматическом режиме;
- проводит анализ качества полученных QSAR-моделей (методом “скользящего контроля”)
- позволяет строить QSAR-уравнения стандартными методами регрессионного анализа “вручную” (множественная линейная регрессия и пошаговая регрессия)

В состав системы входит 23 модуля общим объемом более 1 Мегабайт исполняемого кода. Система BIBIGON эксплуатируется в течение 5 лет в Институте Органической Химии РАН.

4.3.1 Базовые маркеры атомов

В QSAR-системе BIBIGON был реализован следующий вариант базовой маркировки, который успешно прошел практическую проверку (глава 5). Для различения топологических особенностей атомов в графе использовалась степень атома (число ребер у соответствующей вершины М-графа). Был введем соответствующий маркер **p** (p = "power of atom"), который может принимать (для классических органических соединений) семь значений (0, 1, 2, 3, 4, 5, 6). При этом p=0 только для несвязных ("одиночных") атомов в М-графе. Был введен маркер химической связи атома **b**, который опредяется следующим образом:

- **o** - у атома все связи ароматические.
- **a** (aromatic) - у атома есть ароматическая связь;
- **t** (triple) - у атома есть тройная связь;
- **q** - у атома есть три двойные связи;
- **w** - у атома есть две двойных связи;
- **d** (double) - у атома есть двойная связь;
- **s** (single) - все связи атома одинарные;

Метка атома, в которую включены маркеры, записывается в виде строки следующего вида: <имя атома><маркер **p**><маркер **b**>.

Кроме **p** и **b** маркеров используется маркер **r** (ring) - маркер положения атома в кольцевой системе, который опредяется следующим образом:

- **a** (aromatic) - атом “в ароматической системе”;
- **s** (substitute) - атом "кольцевой с заместителем";
- **r** (ring) - атом "чисто кольцевой";
- **l** - атом имеет 2-х ароматических соседей;
- **n** - имеет 1-го ароматичексого соседа;
- **m** - имеет кольцевого соседа;
- **c** (chain) - атом ациклический (цепной).

Таким образом, базовая метка M_i атома v_i имеет вид:

<Имя атома><**p**-маркер><**b**-маркер><**r**-маркер> или "**NNnpbr**",

Имена атома записываются заглавными символами, **p**-маркер - цифрой, **b**- и **r**- маркеры - строчными буквами. Если маркер не используется ("выключен"),

то вместо него в M_i ставится символ "*". Такая запись позволяет не использовать символ разделителя меток атомов и записывать метки подряд друг за другом (в виде конкатенации символов).

4.3.2 Работа пользователя с системой

Работу с системой (после загрузки обучающего множества молекул и выбора свойства) можно условно разделить на следующие три этапа:

- 1-ый этап - преобразование структур "обучающего множества" в векторный вид и формирование матрицы описания.
- 2-ой этап - "обучение" системы: построение QSAR-моделей;
- 3-ий этап - "тестирование" найденных QSAR-моделей.

На вход системы поступает СБД - список структур соединений и список экспериментальных значений свойства этих соединений. На выходе система выдает QSAR-уравнение оценки свойства в виде:

$$Y = b_0 + b_1*F(X_1) + b_2*F(X_2)+...+b_k*F(X_k) + \varepsilon$$

где параметры X_k характеризуют число повторений структурных фрагментов - цепочек маркированных атомов - в исследуемых соединениях, либо числовые топологические индексы.

Первый этап - преобразование структур обучающей СБД. Пользователь с помощью системы проводит:

- диалоговый выбор обучающего множества структур с возможностью поструктурного просмотра;
- диалоговый выбор одного из свойств, его функциональное преобразование (задание зависимой переменной QSAR-моделей), просмотр гистограммы свойства;
- выбор установок и проведение преобразование структур (нахождение ароматических связей, формирование суператомов, "разметка" атомов);
- исчерпывающее перечисление (в преобразованных структурах обучающего множества) цепочек маркированных атомов заданной длины;

- объединение всех найденных цепочек в общий список и формирование QSAR-таблицы обучающего множества;

Общую схему формирования QSAR-таблицы, описывающей молекулы обучающего множества, можно представить следующим образом.

I. Пользователь системы определяет набор маркеров, которые будут участвовать в формировании меток атомов.

II. Система проводит предварительное преобразование М-графов обучающего множества в новые М-графы, имеющие маркированные вершины. При задании установок пользователя новые графы ("обработанные") М-графы" могут содержать:

   A. "суператомы",

   B. атомы водорода как вершины графа ("восстановление водорода"),

   C. разметку ароматических связей в кольцевых системах с чередующимися одинарными и двойными связями.

III. Пользователь задает величину ***KL*** - число атомов в перечисляемых цепочках, система проводим перечисление цепочек, содержащих **1,2, ...** ***KL*** атомов, (в новых, обработанных графах) и подсчитывает число повторений каждой цепочки.

IV. Система объединяет все неэквивалентные цепочки, (заданной длины ***KP***, где ***KL => KP***) в общий список перенумеровываем их и формирует QSAR-таблицу $\boldsymbol{X = (X_{ij})}$, где $\boldsymbol{X_{ij}}$ число повторений j-ой **(*j=1,M*)** цепочки $\boldsymbol{C_j}$ в **i**-ой **(*i=1,N*)** молекуле (обработанном МХ-графе). Столбец $\boldsymbol{X_j}$ QSAR-таблицы соответствует первичному **j**-ому дескриптору - цепочке $\boldsymbol{C_j}$.

Операция системы	*Средства контроля/визуализации*
1. Загрузка обучающей структурной БД.	• Просмотр структурных формул • Просмотр гистограммы распределения значений выбранного свойства
1. Преобразование структур СБД (замена	• Просмотр преобразованных структурных формул

функц. групп атомов на один "суператом").	• Просмотр гистограмм преобразованных значений свойства
1. Нахождение цепочек маркированных атомов (заданной длины).	Просмотр списка цепочек, найденных для каждой молекулы.
1. Построение “Матрицы молекула-фрагмент обучающей СБД.	Просмотр общего списка цепочек атомов.

Второй этап - построение QSAR-моделей. Пользователь с помощью системы проводит:

- ♦ адаптивное формирование "базового списка" признаков (с использованием функциональных преобразований одной переменной);
- ♦ построение линейных QSAR-моделей на "базовых признаках":
  - A. методом пошаговой регрессии;
  - B. методом группового учета аргументов (МГУА-методом);
- ♦ графическая визуализация QSAR-моделей (свойство-модель)

Операция системы	*Средства контроля/визуализации*
1. Отбор “значимых” дескрипторов (с заданным порогом коррелированности)	Просмотр базового списка дескрипторов
1. Пошаговая линейная регрессия (на “значимых” дескрипторах).	• Просмотр QSAR-параметров • Визуализация графика зависимости “свойство-оценка”
1. Комбинационное МГУА-оценивание (на “значимых” дескрипторах).	• Просмотр QSAR-параметров • Графическая визуализация
1. Сохранение QSAR-моделей.	Формирование отчета

Третий этап - "тестирование" найденных QSAR-моделей. Пользователь с помощью системы проводит:

- проведение верификации качества прогностичности построенных QSAR-моделей методом "скользящего контроля";
- визуализация результатов "скользящего контроля";
- прогнозирование свойства новых органических соединений;
- выполнение различных сервисных функций, включая формирование компьютерного отчета о проведении QSAR-моделирования, о полученных параметрах и результатах.

Операция системы	*Средства контроля/визуализации*
1. Нахождение прогностической значимости QSAR-моделей (скользящий контроль).	Формирование отчета
1. Тестирование QSAR-моделей - прогнозирование свойств веществ из “тестовой БД”.	Формирование отчета
1. Построение регрессионного уравнения на дескрипторах, заданных “вручную”.	Формирование отчета

В системе реализован оригинальный графический интерфейс, позволяющий легко анализировать построенные QSAR-уравнения (рисунок 4.16).

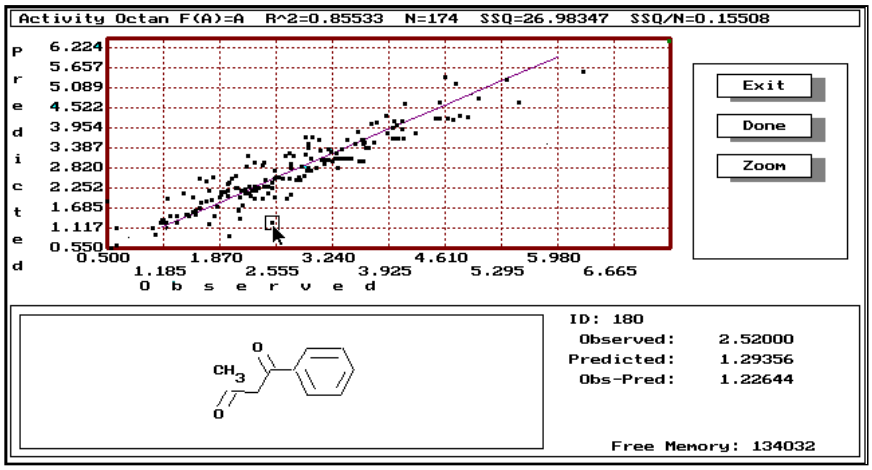


Рисунок 4.16. Пример графического представления в системе BIBIGON построенного QSAR-уравнения.

4.4 Программная система ChemAdd - развитие системы BIBIGON.

QSAR-система BIBIGON представляет собой по сути реализацию экспертного этапа QSAR-анализа, т.е. построение QSAR-моделей и оценку их прогностичного качества. Несмотря на то, что в системе была реализована возможность построения прогноза для новых соединений, большого внимания пользовательскому этапу работы не уделялось.

В новой QSAR-системе ChemAdd, построенной на основе системы BIBIGON, основные модификации и расширения коснулись в основном пользовательского этапа работы. Для решения “обратной задачи” (поиск новых соединений, обладающих заданными свойствами - см. 6 главу) в систему введен графический редактор-генератор М-графов, подробно рассмотренный в 6 главе. В экспертную часть системы включен редактор обучающей выборки, позволяющий в графическом виде ввести набор М-графов с приписанными им экспериментальными значениями свойств. Введенные выборки затем используются для построения QSAR-моделей.

4.4.1 Редактор обучающей выборки химических структур.

Модуль написан на основе объектно-ориентированной библиотеки, рассмотренной выше. Пользовательский интерфейс реализован с использованием стандартной библиотеки **Turbo Vision,** которая была модифицирована для работы с графикой. Общий размер программных текстов 41000 строк (включая используемые бибиотеки), размер EXE-модуля 300 Кб. Программа позволяет:

- Использовать выборки химических структур (базы данных) в двух коммуникационных текстовых форматах
- Создавать новые и модифицировать старые базы данных, используя редактор полей (рис. 4.17).
- Добавлять новые и модифицировать старые структуры в базу данных одновременно с вводом и редактированием данных о свойствах (рис. 4.18)

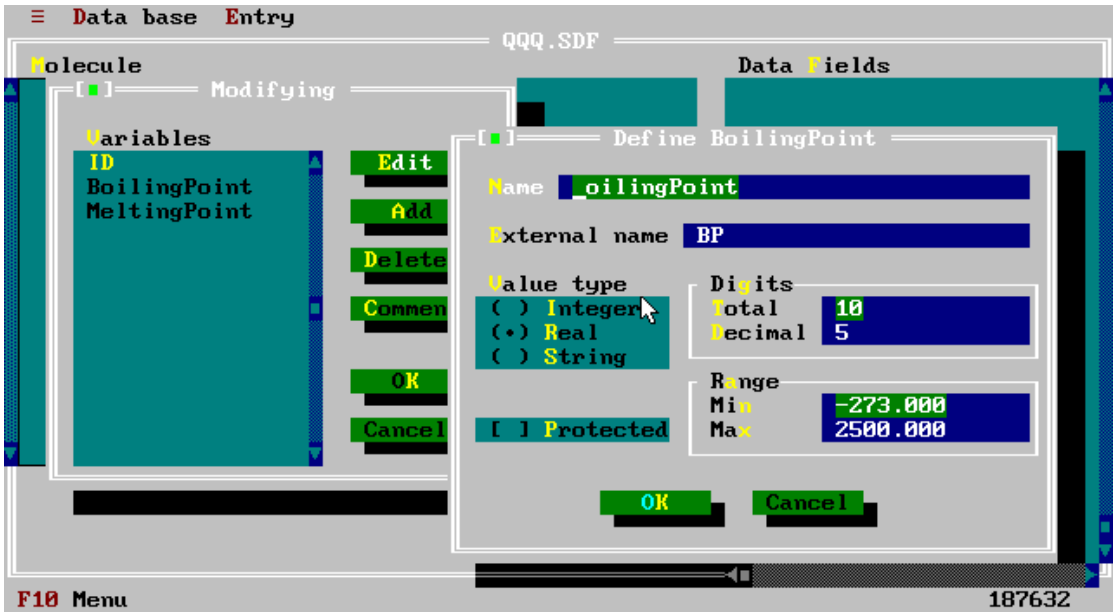


Рисунок 4.15 Определение полей в Редакторе обучающей БД.

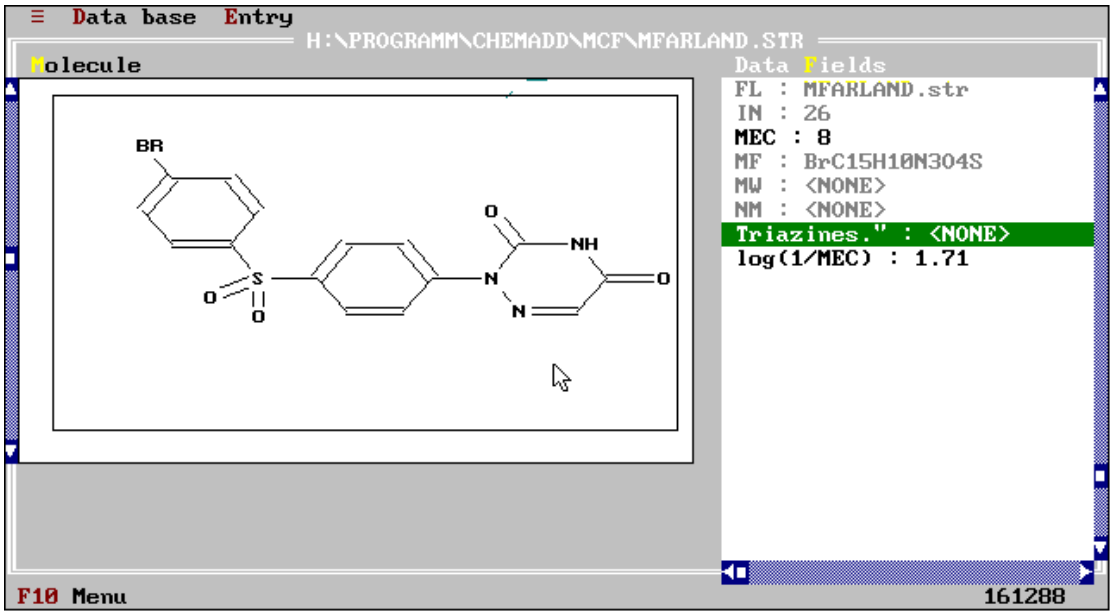


Рисунок 4.16. Общий вид главного окна редактора обучающей СБД.

В процессе работы над модулем был создан интерфейсный объект - окно графического редактора молекулярных структур (М-графов), который был затем использован для построения модуля редактора-генератора М-графов (рис.4.19).

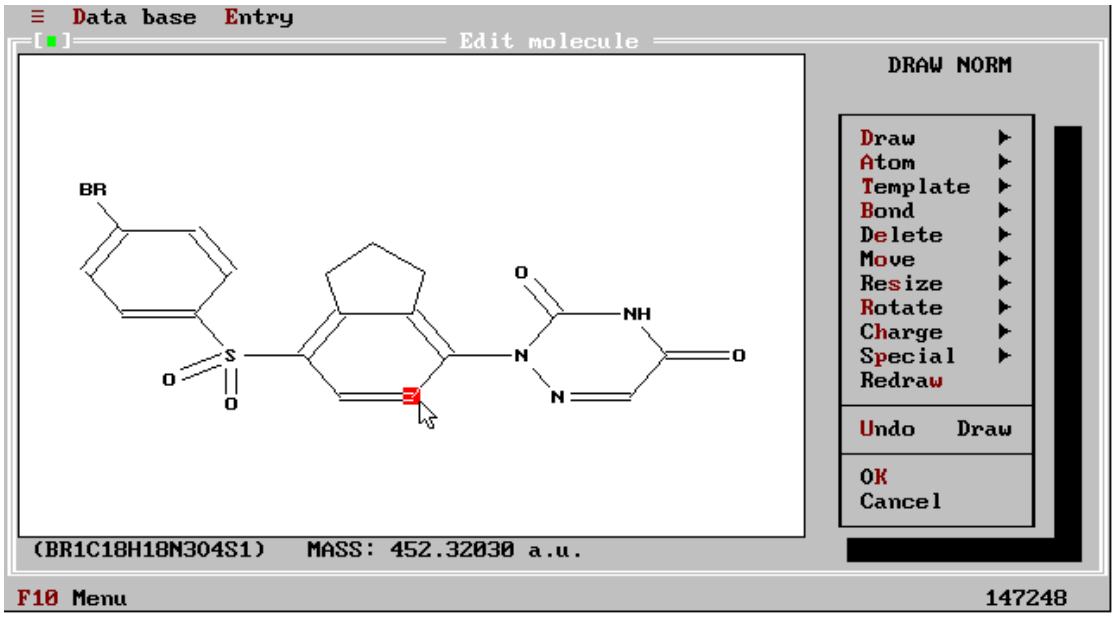


Рисунок 4.17 Общий вид окна графического редактора М-графов.

Редактор М-графов позволяет проводить все необъходимые операции для визуального создания и редактирования химической структуры:

- добавляит/удалять атомы (вершины М-графа)
- изменять название и заряд атома
- добавлять/удалять связи (ребра М-графа)
- изменять кратность связей
- перемещать атомы и связи
- изменять размер структуры
- вращать структуру
- проверять корректность расстановки связей (контролировать валентность атомов)
- динамически во время рисования вычислять брутто-формулу и массу молекулы

4.4.2 Массовая генерация и селекции графов как метод решения обратной QSAR-задачи.

Химика, занимающегося прикладными разработками, интересует прежде всего вид новых химических структур, которые обладают свойствами в заданном интервале значений. QSAR-уравнения являются для него вспомогательным средством для конструирования новых структур. Химику необходимо найти решение так называемой "обратной задачи": т.е. по QSAR-

уравнениям найти структуры, оценка свойств которых лежит в заданном интервале.

Тривиальное решение этой задачи может быть получено в результате "ручного подбора": пользователь вводит в компьютер структуру нового соединения, а система выдает значение оценки С свойства. С практической точки зрения здесь возникает следующая проблема. Химиков интересует, как правило, не одно свойство, а одновременно несколько свойств (спектр свойств). Например, вещество должно обладать не только лекарственным действием (свойство 1), но и одновременно малой токсичностью (свойство 2) и высокой скоростью вывода из организма (свойство 3). В связи с оценкой спектра свойств "ручной" подбор молекул-кандидатов для органического синтеза представляется трудоемкой и бесперспективной процедурой. Возникает задача генерирования ЭВМ различных молекулярных структур, которые затем будут оценены отдельно по каждому исследуемому свойству.

Метод массовой генерации М-графов, реализованный в системе “ChemAdd” заключается в следующем. Используя редактор-генератор химических структур, пользователь рисует так называемую "граф-основу" и указывает к каким местам этой основы могут быть присоединены заместители (группы атомов) (рис. 4.20). Затем задается список этих заместителей для каждого из указанного мест. После запуска программа начинает генерировать все возможные сочетания заместителей из заданных списков, присоединяя их к основе. В списках заместителей могут рекурсивно присутствовать генерируемые группы. Это предоставляет пользователю мощное средство по генерации большого количества разнообразных соединений при весьма простом входном графическом языке описания задачи генерации. Во время работы генератор автоматически отслеживает появление изоморфных химических графов.

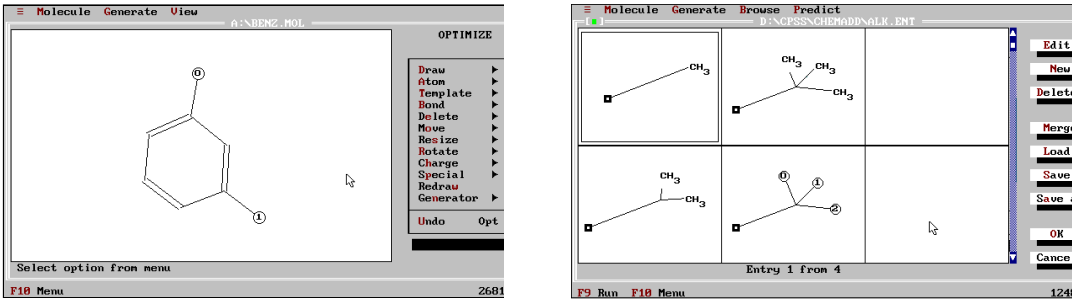


Рисунок 4.18. Пример задания в редакторе-генераторе основы и списка заместителей для генерации М-графов.

Результат работы генератора является список М-графов, который подается на вход "прогнозных фильтров", задаваемых пользователем (рис. 4.21.). Каждый прогнозный фильтр включает в себя семейство QSAR-моделей и допустимый диапазон оценки свойства. Сгенерированная химическая структура проходит через фильтр (на дальнейшую обработку) только в том случае, если предсказанное значение ее свойства лежит в заданном диапазоне.

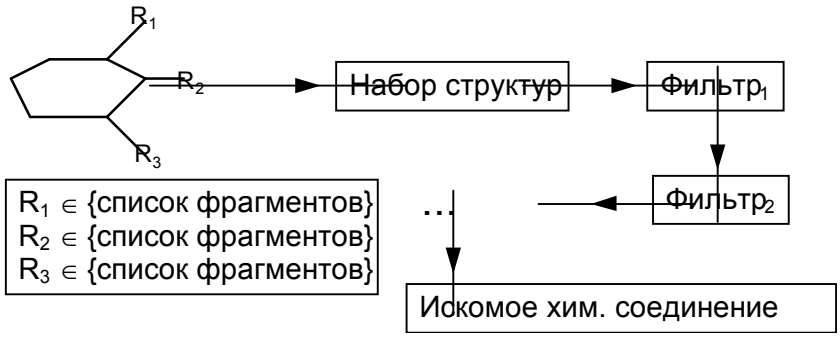


Рисунок 4.19. Схема решения “обратной QSAR-задачи” на основе генератора структур типа “основа-заместители”.

4.4.3 Проблема отказа от прогноза свойства сгенерированной структуры.

В результате массовой генерации создается большое число М-графов, часто принадлежащих к таким химическим классам, для которых используемая в качестве фильтра QSAR-модель неприменима. Возникает задача поиска условий отказа от таких “плохих” структур. Применение для построения QSAR-моделей эволюционного МГУА-алгоритма позволяет решить эту проблему следующим образом.

В МГУА-процессе на каждом шаге вычисления формируется *семейство QSAR-уравнений*, описывающих одно и то же свойство на основе различных наборов дескрипторов. Вектор ($S_1,S_2,...S_p$) оценки свойства по нескольким уравнениям, входящим в QSAR-модель (где S_i - оценка i-го уравнения, $i=1,...,p$), используется для получения значения оцениваемого свойства (S_0) и оценки стандартной ошибки вычисления (*err*). Окончательная оценка S_0 вычисляется путем усреднения значений оценок всех QSAR-уравнений: $S_0 = \frac{\sum_{i=1}^{p} S_i}{p}$, и величина ошибки находится как $err = 2 \cdot \sqrt{D(S)}$, где *D(S)* - оценка дисперсии S_i: $D(S) = \frac{\sum_{i=1}^{p} S_i^2}{p} - \left(\frac{\sum_{i=1}^{p} S_i}{p} \right)^2$ (рис. 4.22).

Для вычисления S_0 может использоваться “метод голосования”, при котором S_0 вычисляется как взвешенная сумма $S_0 = \sum_{i=1}^{n} w_i \cdot S_i$, $\sum_{i=1}^{n} w_i = 1$. В этом случае веса w_i вычислены заранее на “экспертном” этапе и должны храниться вместе с QSAR-уравнениями. Для окончательного отбора “плохих” М-графов используются комбинации следующих правил:

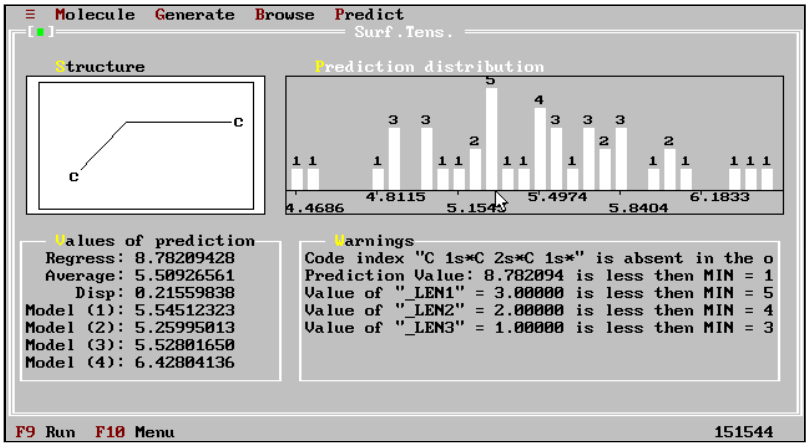


Рисунок 4.20. Результат коллективного прогноза по QSAR-модели для отдельного соединения.

1. Пусть задано пороговое значение ошибки оценки свойства (T_{err}). QSAR-модель считается неприменимой к М-графу, если вычисленное значение *err* ошибки превосходит порог: $err > T_{err}$.
2. Если QSAR-модель была построена на определенном классе соединений, содержащих специфические химические фрагменты, то ее предсказательную возможность ограничим именно этим классом: определяются так называемые “индикаторные фрагменты”, т.е. фрагменты, присутствующие во всех соединениях обучающей выборки. Предсказание не проводится, когда входной М-граф не содержит всех “индикаторных фрагментов”.
3. На М-графах обучающей выборки вычисляется ряд “глобальных” теоретико-графовых инвариантов ($I_1, I_2, .., I_n$). Для каждого инварианта I_k определяется минимальное и максимальное значения по всей обучающей выборке (I_k^{min} и I_k^{max} соответственно). Для входного М-графа перед прогнозированием рассчитываются значения этих инвариантов, и для каждого инварианта I_k проверяется попадание его в интервал $[I_k^{min}-\Delta_k, I_k^{max}+\Delta_k]$. Предсказание не проводится, если $\exists\, I_k: I_k \not\subset [I_k^{min}, I_k^{max}]$.

4.5 Организация базы знаний, надстраиваемой над структурными БД.

Извлечение и использование знаний о зависимостях свойств химических веществ от их структуры (количественных корреляций “структура-свойство” - QSAR) может выявлять противоречия и ошибки в уже введенных данных, а также производить постоянный контроль непротиворечивости при вводе новых данных.

Для каждого из свойств, представленных в БДХВ, формируется QSAR-модель. Свойства большинства соединений будут описываться моделью удовлетворительно. Среди оставшихся соединений (с помощью эксперта-химика) можно выделить структуры, для которых построенная QSAR-модель недостаточно точна, и, с помощью введения дополнительных условий, доопределить модель. Соединения, которые продолжают "выпадать" из доопределенных моделей (т.е. иметь большую ошибку аппроксимации свойства по QSAR-модели), требуют внимательного рассмотрения на предмет существования ошибки или в структуре графа, или в значении введенных

свойств. ¶ Созданные модели сохраняются в виде алгоритмической базы знаний о свойствах, образуя интеллектуальную надстройку над существующей БДХВ.

Наличие базы знаний о свойствах химических структур позволяет придать универсальный характер БДХВ для характеристики определенных химических классов. Действительно, занесение в структурную базу всех данных об известном классе химических соединениях является трудной задачей, при этом практически невозможно гарантировать “полный охват” даже для ограниченного класса веществ. С другой стороны, пользователю достаточно иметь только оценку свойства для интересующего его соединения. Единообразное представление QSAR-моделей в виде алгоритмической базы знаний, надстраиваемой над существующими БДХВ, позволяет проводить оценки свойств всех соединений данного химического класса.

Для практического применения построенных QSAR-моделей в решении обратной задачи и для использования совместо с фактографическими базами данных требуется обеспечить их сохранение в некотором стандартном виде. Наиболее рациональным решением представляется организация QSAR-моделей в виде QSAR-базы знаний.

База знаний содержит факты (данные) и правила для принятия решения на основе работы “механизма вывода” [333]. В каждой системе, основанной на знаниях, структура и способы представления знаний существенно зависят от специфики предметной области. Принципиальным моментом в построении таких систем является отделение знаний о предметной области от других знаний системы и выделение их в особую базу данных (базу знаний), которая используется “механизмом вывода” (решателем задач системы) для принятия решения. Внесение тех или иных изменений в базу знаний системы должно влиять на принятие решений “механизмом вывода” без каких-либо изменений в программных компонентах системы. Особенности организации базы знаний в этом случае состоят в следующем:

1.Запросом к системе является М-граф, т.е. структурная формула молекулы.

2.Ответом системы является оценка свойства молекулы или отказ.

3.Базой фактов являются библиотеки фрагментов с весовыми коэффициентами для оценки свойства.

4.Правила в базе знаний представлены (для каждой библиотеки фрагментов) условиями для маркирования структурных примитивов в М-графе - запросе, с одной стороны. И условиями для кодирования и сравнений (с фрагментами библиотек БЗ) фрагментов, присутствующих в М-графе - запросе, с другой.

5.Механизм вывода состоит в последовательном применении условий, хранящихся в БЗ, к М-графу - запросу.

QSAR-база знаний системы прогнозирования содержит следующую информацию по каждой QSAR-модели (рис.4.23):

- ♦список QSAR-уравнений для оценки свойства,
- ♦библиотека структурных фрагментов,
- ♦таблицы для вычисления окончательной оценки,
- ♦правила вычисления ошибки оценки свойства,
- ♦правила фильтрации входных М-графов.

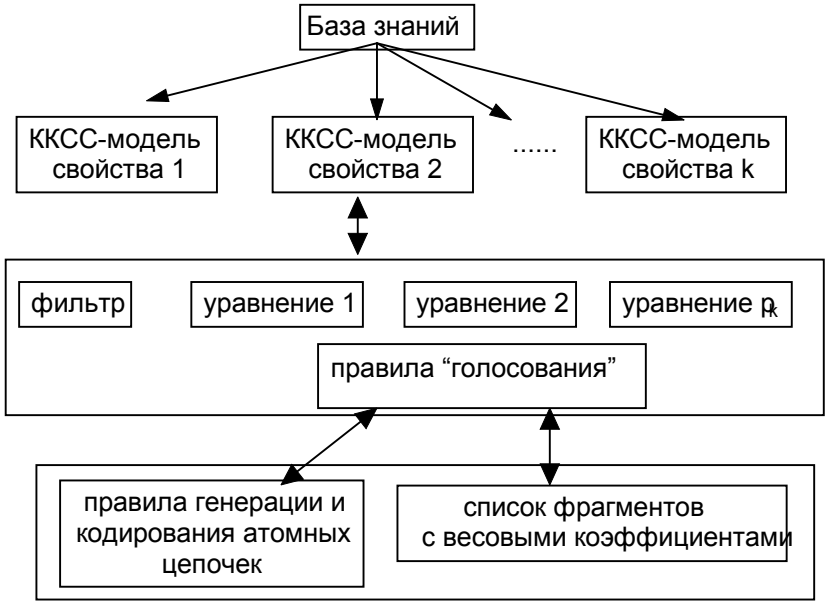


Рисунок 4.21. Организация QSAR-базы знаний.

Описанный подход к формированию QSAR-базы знаний над фактографической базой данных был реализован практически для двух больших баз данных: для базы данных по липофильности (LogP) органических

соединений, содержащей более 3500 структур, и для базы данных по ^{13}C NMR спектрам, содержащей более 20000 структур [334].

4.6 Выводы

Рассмотрена модульная организация программной системы для построения QSAR-моделей на основе принципов, изложенных во второй и третьей главах работы. Описана внутренняя организация структур данных системы BIBIGON и основные этапы работы пользователя. Изложены отличия системы ChemAdd, которая представляет собой развитие системы BIBIGON. Отмечено, чтоосновной акцент в ней сделан на организацию выполнения пользовательского этапа. Для решения “обратной QSAR-задачи” в систему ChemAdd включен графический редактор-генератор М-графов. В экспертную часть системы ChemAdd включен редактор обучающей выборки, позволяющий в графическом виде вводить М-графы вместе со значениями их свойства.

Рассмотрено решение обратной QSAR-задачи на основе массовой генерации и селекции М-графов, который заключается в использовании фрагмента-"основы" и комбинаторного перебора фрагментов-заместителей. Результатом работы генератора является список М-графов, который подается на вход "прогнозных фильтров", задаваемых пользователем. Каждый прогнозный фильтр включает в себя МГУА-семейство QSAR-моделей и допустимый диапазон оценки свойства. Химическая структура проходит через фильтр только в том случае, если предсказанное значение ее свойства лежит в заданном диапазоне.

Описано представление QSAR-моделей в виде базы знаний и обсуждены особенности ее организации.

5. Глава 5. Прогнозирование физико-химических свойств веществ.

В главе приведены практические результаты, полученные при использовании созданных программных систем и иллюстрирующие эффективность методики автоматического построения и выбора адекватного описания структур для **QSAR**-задач. В ходе выполнения работы было построено *более трехсот* различных **QSAR**-моделей. Ниже приводятся лучшие из полученных моделей вместе с результатами их проверки на прогностичность методом "скользящего контроля". Большинство обучающих СБД были введены в компьютер с помощью структурного редактора МОЛЕД, разработанного на Химфаке МГУ [335].

Физико-химические свойства веществ, как правило, хорошо прогнозируются с помощью аддитивных схем расчета. Ставилась цель автоматически построить аналоги таких схем с использование разработанного комплекса программ на основе связных k-фрагментов. Исследовались различные варианты “базовой” маркировки атомов и ее влияние на прогностичность получаемых **QSAR**-моделей. Для каждого свойства описаны сводные результаты вычислительных экспериментов и их сравнение с литературными данными. Выбор признакового пространства осуществляется путем многократных вычислительных экспериментов и верификаций **QSAR**-уравнений, построенных на структурных дескрипторах разной степени сложности. Проводился переход от простых базовых фрагментов к сложным и для каждого признакового пространства находились (методом скользящего контроля) и сохранялись прогностически устойчивые **QSAR**-модели. Лучшие из таких моделей могут быть использованы для предсказания свойств новых структур. Результаты проведенных расчетов, полученные при использовании созданной программной **QSAR**-системы, иллюстрируют эффективность предложенной методики автоматического выбора адекватного описания структур на основе использования цепочек маркированных атомов для построения **QSAR** моделей[336].

Для всех **QSAR** моделей приведены статистические параметры, полученные методом "*скользящего контроля*" (Cross-validation) следующим образом: из обучающей СБД удалялось i-ое соединение и на данном наборе дескрипторов строилась **QSAR**-модель. По найденной модели проводился

расчет-предсказание свойства удаленного i-ого соединения, а ошибка предсказания запоминалась. Значение i пробегает все N структур обучающей СБД, и формируется "прогнозный" вектор ошибки PRESS (PREdict Sum of Squares). По вектору PRESS вычислялись соответствующие параметры модели: R^2_{cv}, S, F - квадрат множественного коэффициента регресии, стандартное отклонение и критерий Фишера.

Коэффициенты **QSAR**-уравнений приведены в виде:

4. -0.443 (0.021) (W: 67) F(X)=X H 1C 4**C 4**C 4**H 1****

где **"4."** - номер переменной;

"-0.443" - значение весового коэффициента $\mathbf{b_j}$;

"(0.021)" - отклонение весового коэффициента $\mathbf{b_j}$;

"(W: 67)" - значение суммы столбца дескриптора; позволяет контролировать "вырожденные признаки, для которого имеем **"(W: 1)"** или **"(W: 2)"**

"F(X)=X" - указывает, какое функциональное преобразование было выполнено над вектором дескриптора (если **"F(X)=X"** - то преобразование значения дескриптора не проводилось);

"H 1C 4**C 4**C4**H 1**"** - код цепочки **"H-C-C-C-H"** с **p**-маркером (**b**-маркер и **r**-маркер "выключены" и на их месте присутствует символ "*").

Код цепочки является "самоописывающимся", т.е. по его виду можно полностью восстановить вид цепочки атомов (и топологические и химические особенности атомов по соответствующим маркерам). Параметр **"_LENk" (k=1,2,3,...)** обозначает общее число цепочек длины **k**. При проведении расчетов в некоторых моделях атом водорода **"H"** рассматривался как вершина М-графа и включался в цепочки маркированных атомов. Для этого предусмотрена специальная опция в программе BIBIGON.

5.1 Свойства алканов.

Оценка свойств алканов выполнена как тестовый расчет. Использовано обучающая СБД из 74-х соединений. Выборку ввел в ЭВМ Девдариани Р.О. (Химфак МГУ) на основе работы [337]. Оценивались: температура кипения, молекулярный объем, молекулярная рефракция, энтропия испарения, критическая температура, критическое давление, поверхностное натяжение и температура плавления. Приведены результаты только "скользящего

контроля" за исключением температуры плавления - "нелинейного" свойства, для которого получена плохая прогностичность. Результаты расчетов объединены в Таблица 5.1.

Таблица 5.1 ***Результаты (Скользящий Контроль) QSAR-моделей свойств алканов***.

Свойство	K	R^2_{cv}	S	F
Темп. кип.	3	0.982	6.29	1291.79
Мол.объем.	4	0.998	0.683	10823
Мол. рефр.	2	0.9998	0.058	271389
Энтр. исп.	5	0.9994	0.176	12562
Крит.темп.	3	0.959	11.74	415
Крит.давл.	4	0.946	1.03	307
Пов.натяж.	6	0.969	0.352	324
Темп.плав.	7	574	23.67	1.21

Температура кипения.

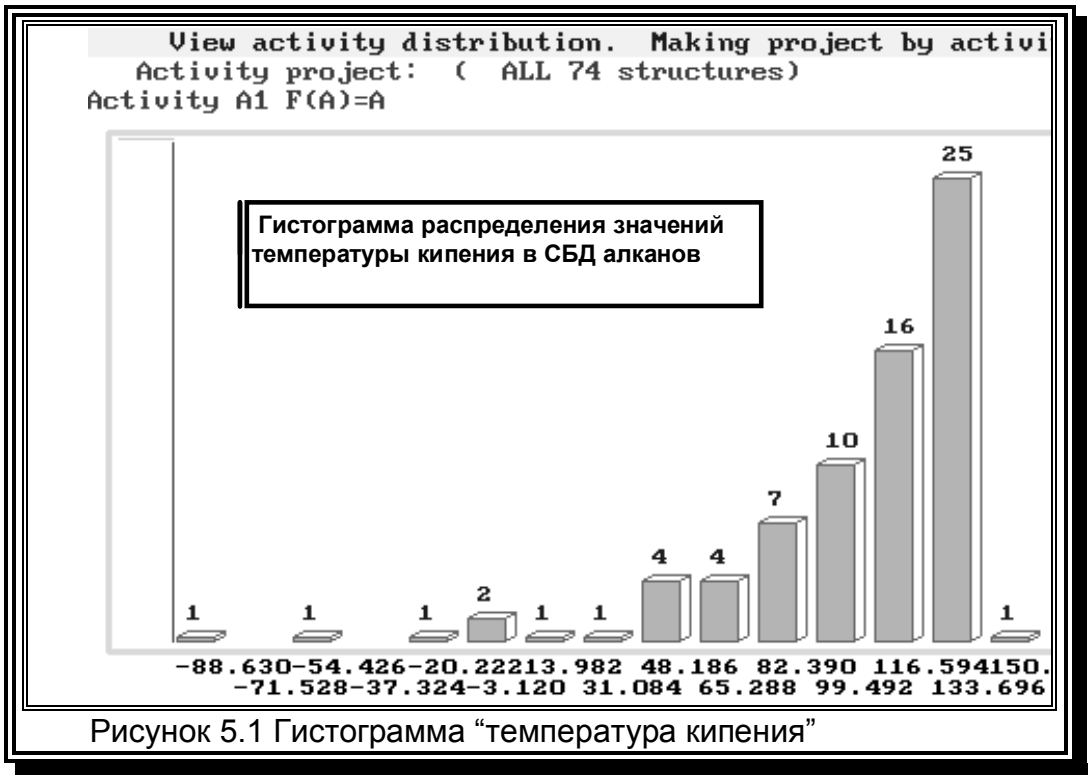


Рисунок 5.1 Гистограмма “температура кипения”

Гистограмма свойства “температура кипения” имеет следующий вид:

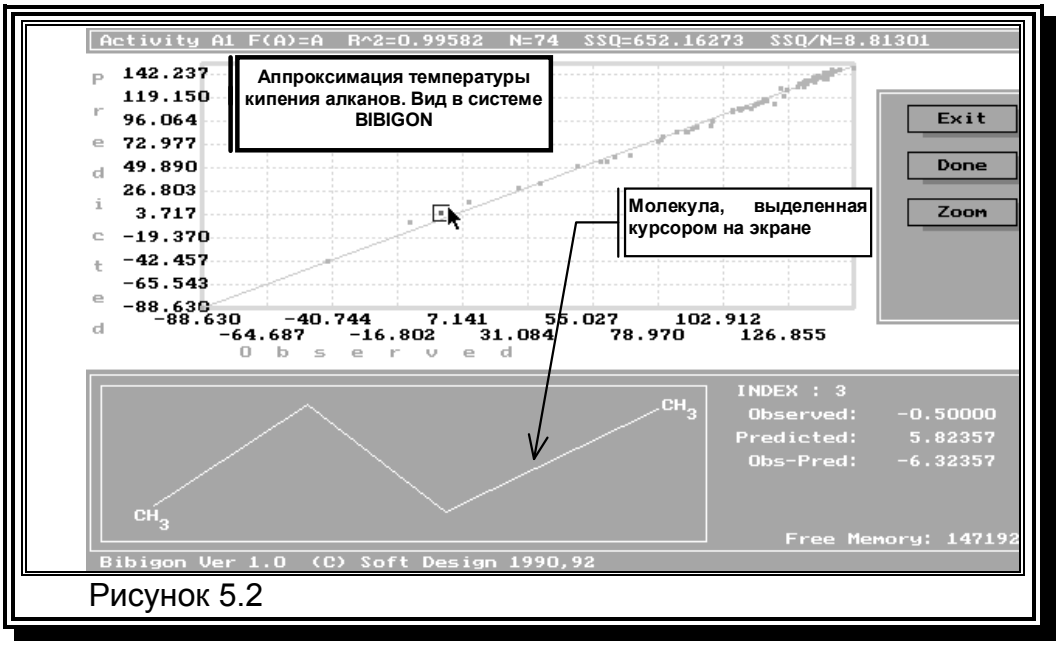


Рисунок 5.2

Скользящий Контроль: **N = 74, K = 3, R^2 = 0.982, S = 6.298, F = 1291.79**

Коэффициенты **QSAR**-уравнения:

1. -132.70 свободный член
2. 33.859 цепочка F(X)=X C 4**
3. -2.272 цепочка F(X)=X C 4**C 4**C 4**C 4**C 4**
4. -0.443 цепочка F(X)=X H 1**C 4**C 4**C 4**H 1**

Молекулярный объем

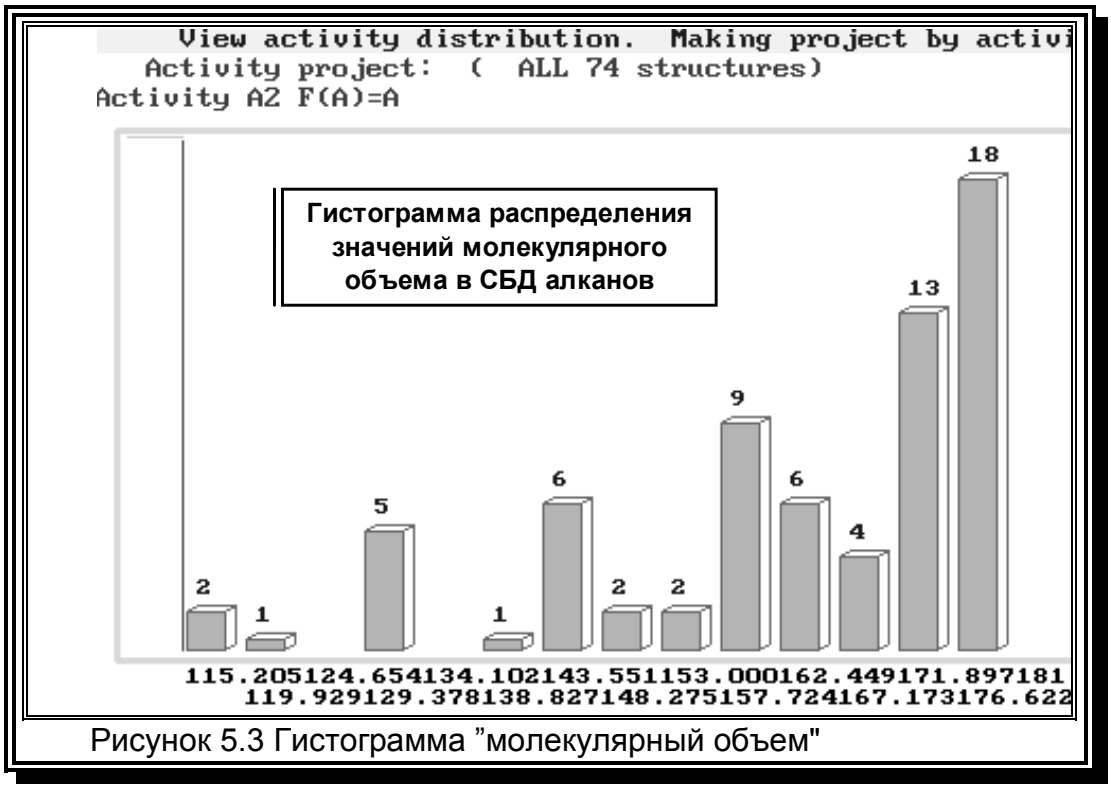


Рисунок 5.3 Гистограмма ”молекулярный объем"

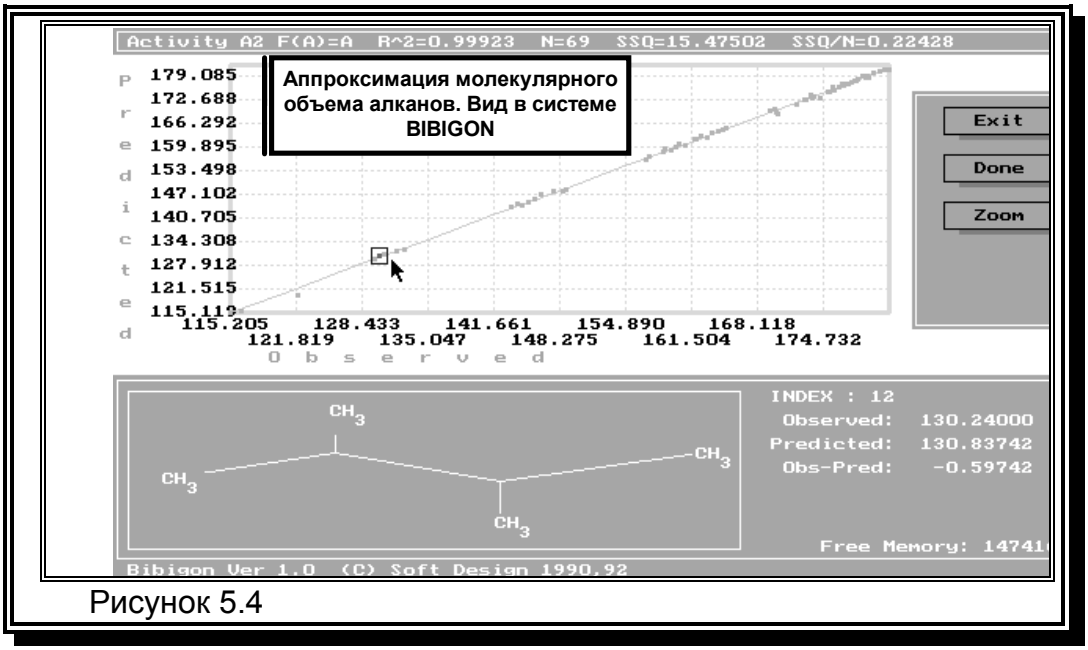


Рисунок 5.4

Гистограмма свойства ”молекулярный объем” имеет следующий вид:
Скользящий Контроль: **N = 69, K = 4, R^2_{CV} = 0.998, S = 0.683, F = 10823**
Коэффициенты **QSAR**-модели:

1. 47.759 (0.169) CONST
2. -0.874 (0.035) (W: 21) C 1scC 2scC 2scC 2sc
3. -0.440 (0.107) (W: 6) C 1scC 4scC 2scC 4sc
4. 18.056 (0.023) (W: 490) _LEN2
5. -1.879 (0.012) (W: 479) _LEN4

Молекулярная рефракция.

Гистограмма свойства ”молекулярная рефракции” имеет следующий вид:

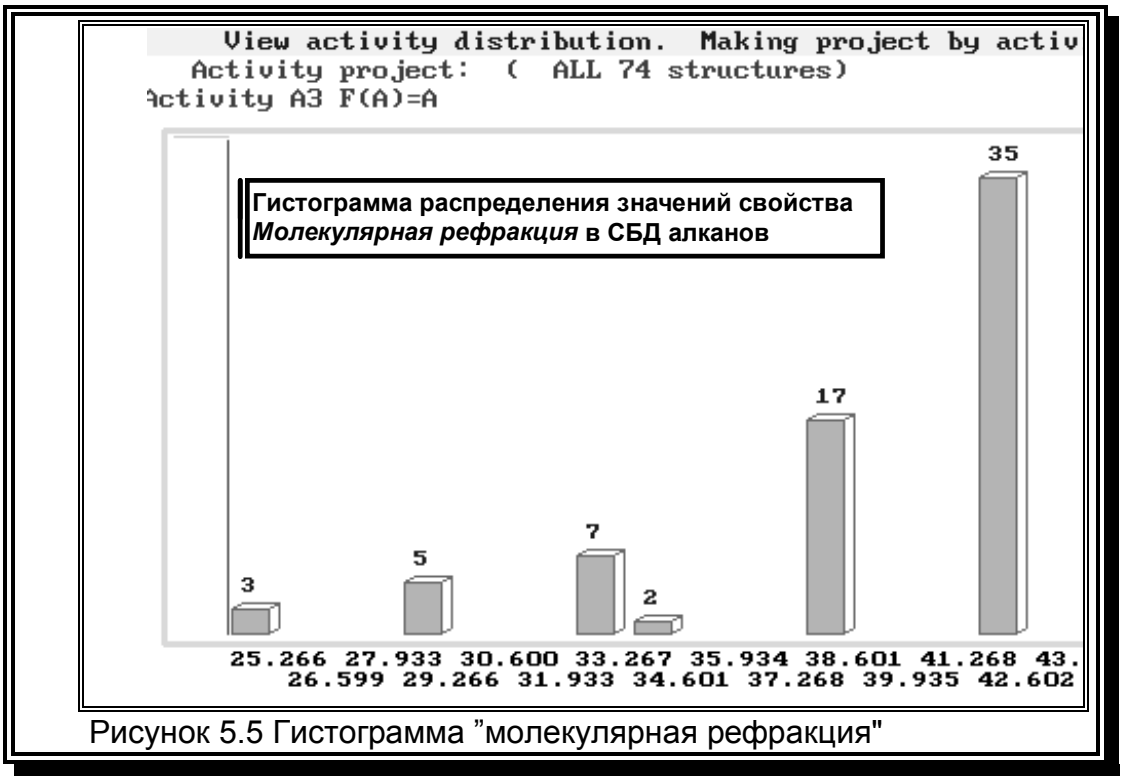


Рисунок 5.5 Гистограмма ”молекулярная рефракция"

Рисунок 5.6

"Скользящий Контроль": **N = 69,K=2, R^2_{CV} = 0.9998, S = 0.058, F = 271389**

Коэффициенты **QSAR**-модели:

1. **1.668 (0.014) CONST**
2. **4.776 (0.002) (W: 559) Chain F(X)= X _LEN1**
3. **-0.126 (0.001) (W: 479) Chain F(X)= X _LEN4**

Энтальпия испарения.

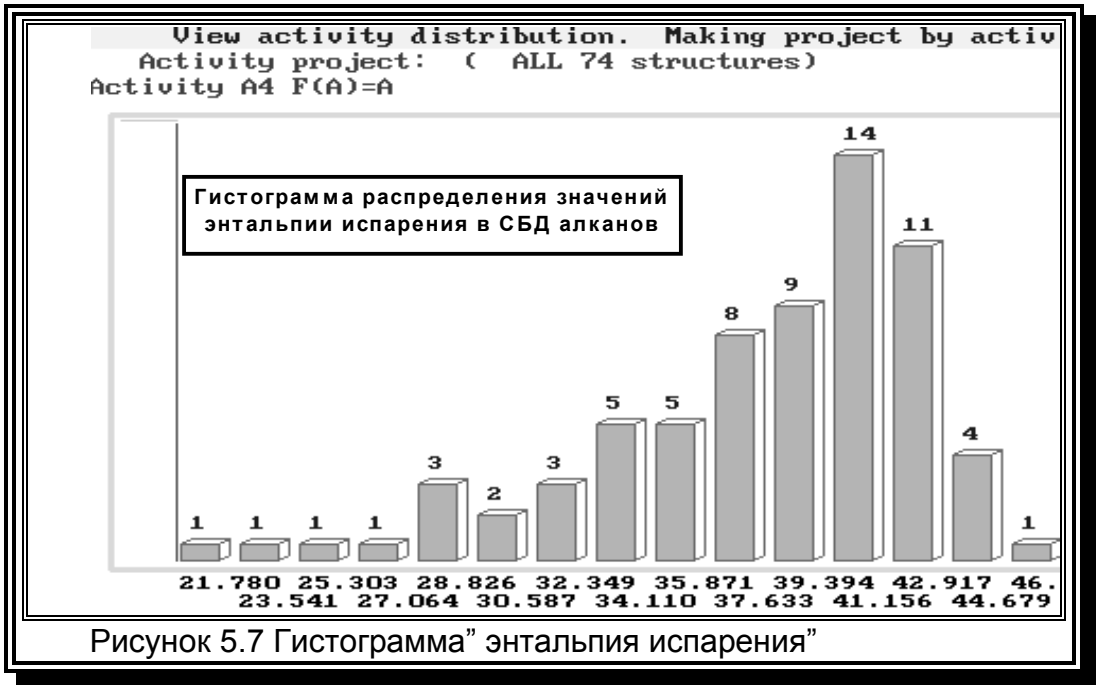


Рисунок 5.7 Гистограмма" энтальпия испарения"

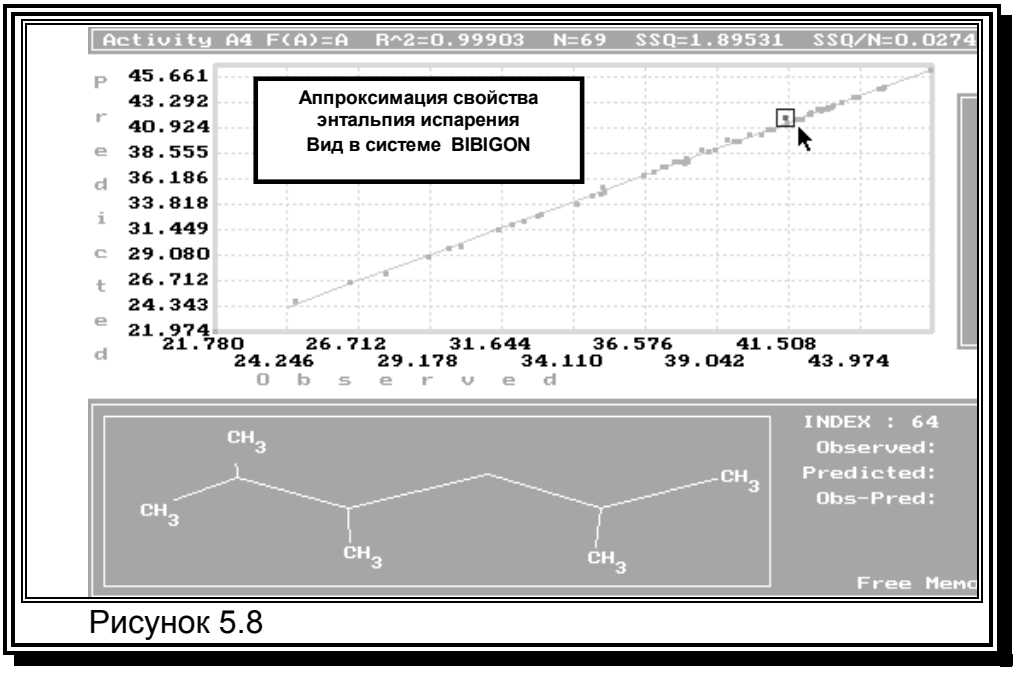


Рисунок 5.8

Гистограмма свойства " энтальпия испарения" имеет следующий вид:
"Скользящий Контроль": **N = 69, K = 5, R^2_{CV} = 0.9994, S = 0.176, F = 12562**

Коэффициенты **QSAR**-модели:

1. **2.811 (0.209) CONST**
2. **0.415 (0.024) (W: 71) Chain F(X)= X C 2scC 2sc**
3. **-0.087 (0.026) (W: 18) Chain F(X)= X C 2scC 4scC 2sc**
4. **5.022 (0.050) (W: 559) Chain F(X)= X _LEN1**
5. **-0.993 (0.022) (W: 588) Chain F(X)= X _LEN3**
6. **0.481 (0.017) (W: 479) Chain F(X)= X _LEN4**

Критическая температура.

Гистограмма свойства “критическая температура” имеет следующий вид:
Скользящий Контроль: **N = 74, K = 3, R^2_{CV} = 0.959, S = 11.74, F = 415**

Рисунок 5.10 Гистограмма “критическая температура"

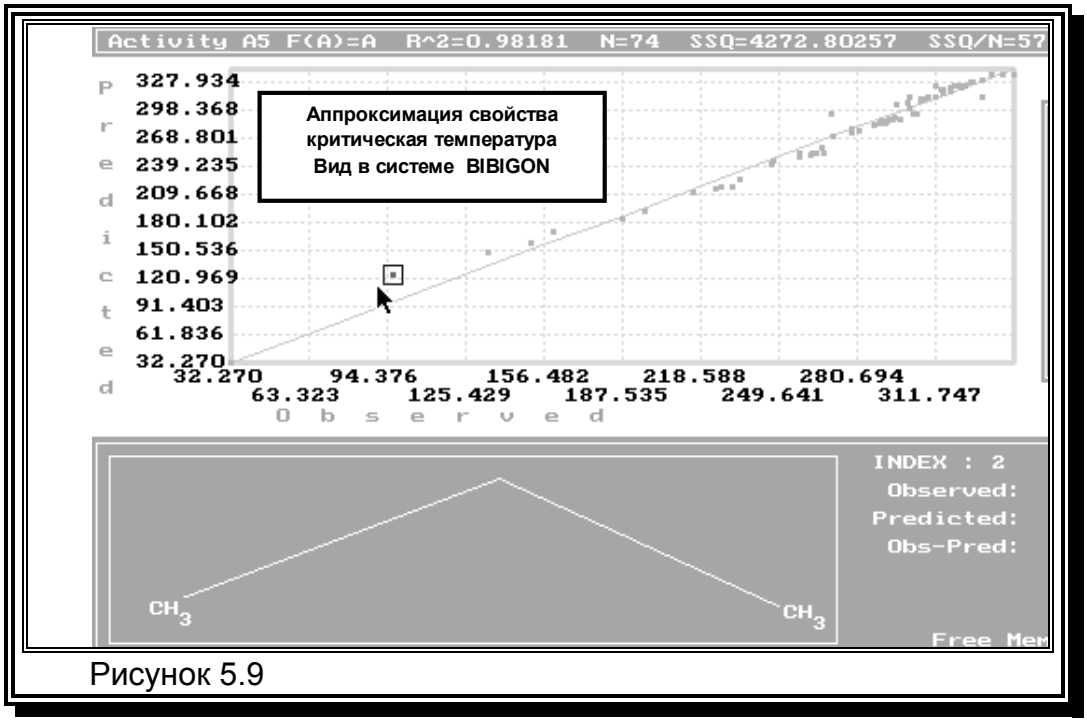


Рисунок 5.9

Коэффициенты **QSAR**-модели:

1. **40.305 (3.1646) CONST**
2. **-67.506 (15.970) (W: 1) Chain F(X)= X C 1scC 1sc**
3. **30.197 (0.472) (W: 580) Chain F(X)= X _LEN1**
4. **-3.935 (0.196) (W: 606) Chain F(X)= X _LEN3**
5. **5.602 (0.155) (W: 489) Chain F(X)= X _LEN4**

Критическое давление.

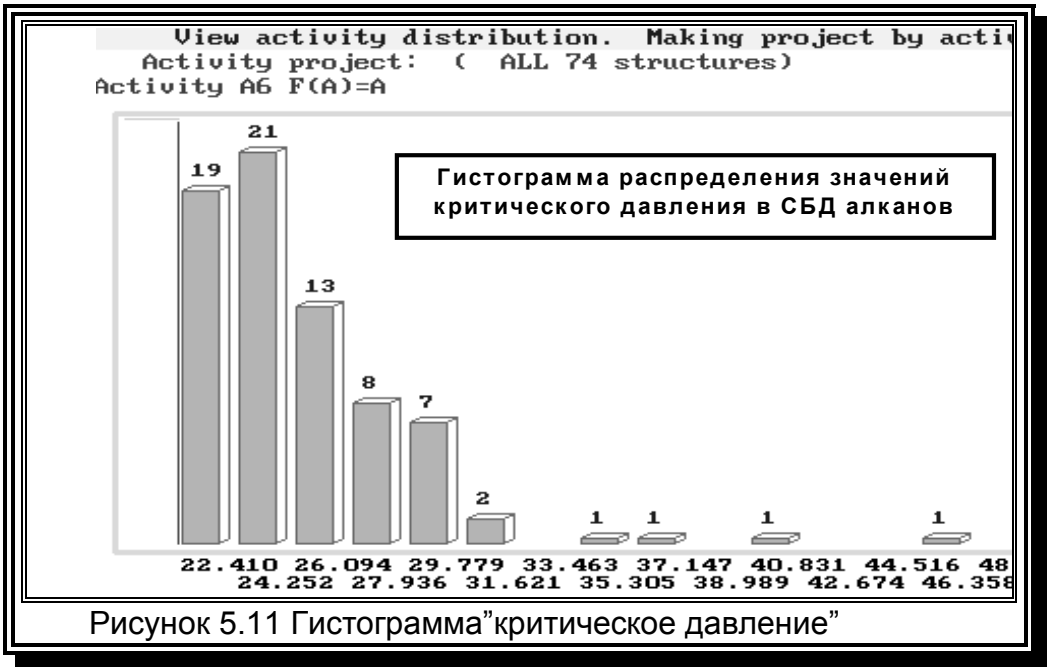


Рисунок 5.11 Гистограмма"критическое давление"

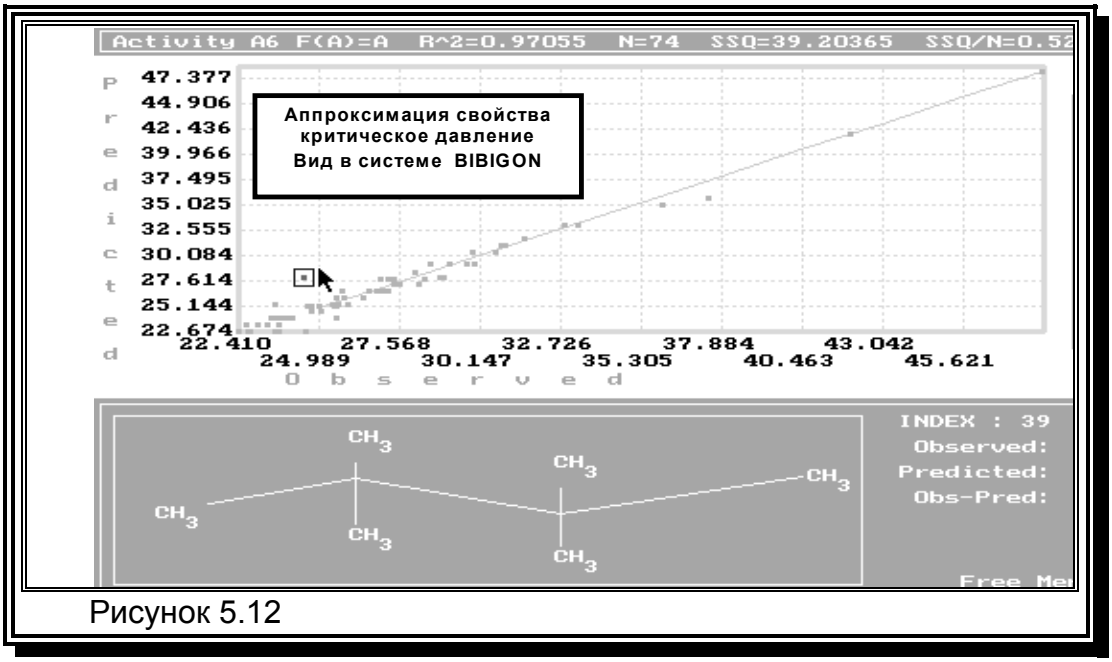


Рисунок 5.12

Гистограмма свойства "критическое давление" имеет следующий вид:
Скользящий Контроль: **N = 74, K = 4, R^2_{CV} = 0.946, S = 1.03, F = 307**

Коэффициенты **QSAR**-модели:

1. **49.433 (0.482) CONST**
2. **0.189 (0.0127) (W: 63) Chain F(X)= X C 1scC 4scC 2sc**
3. **-4.323 (0.110) (W: 506) Chain F(X)= X _LEN2**
4. **0.850 (0.024) (W: 489) Chain F(X)= X _LEN4**
5. **0.342 (0.040) (W: 304) Chain F(X)= X _LEN5**

Поверхностное натяжение.

Гистограмма свойства ”поверхностное натяжение” имеет следующий вид:

Рисунок 5.14 Гистограмма”поверхностное натяжение”

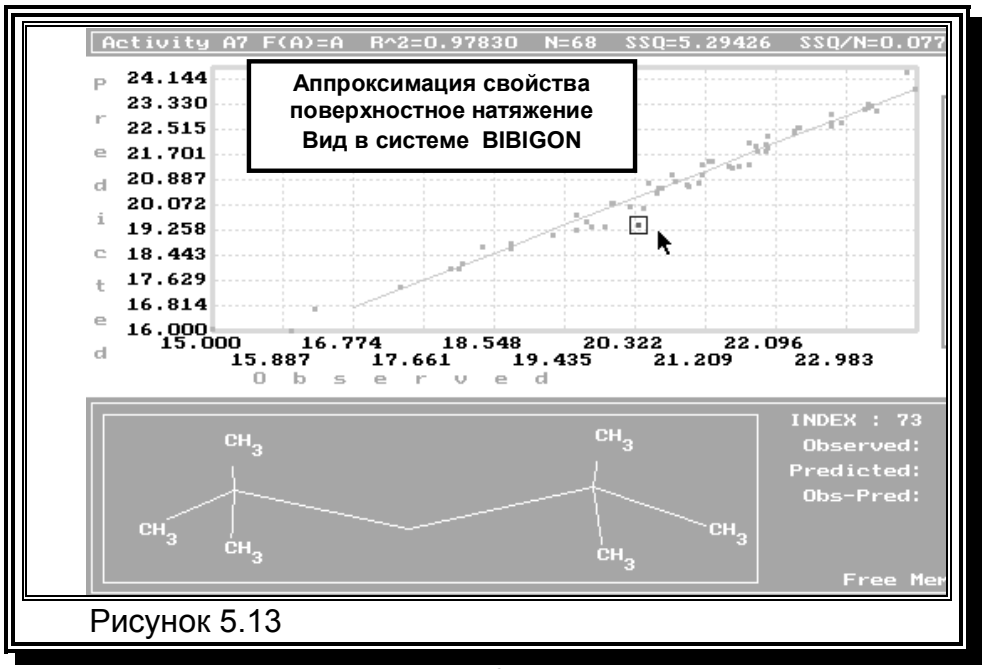


Рисунок 5.13

Скользящий Контроль: **N = 68, K = 6, R^2_{CV} = 0.969, S = 0.352, F =324**

Коэффициенты **QSAR**-модели:

1. 12.257044 (0.152525) CONST
2. 0.431729 (0.010212) (W: 177) Chain \ 93 F(X)= X C 2sc
3. 0.208964 (0.038664) (W: 7) Chain \ 97 F(X)= X C 2scC 2scC 2scC 2scC 2sc
4. -0.160296 (0.015088) (W: 35) Chain \ 127 F(X)= X C 2scC 4sc
5. 0.299550 (0.016567) (W: 17) Chain \ 139 F(X)= X C 3scC 3sc
6. 0.415459 (0.025263) (W: 554) Chain \ 147 F(X)= X _LEN1
7. 0.590131 (0.009415) (W: 479) Chain \ 150 F(X)= X _LEN4

Температура плавления.

Гистограмма свойства ” температура плавления” имеет следующий вид:

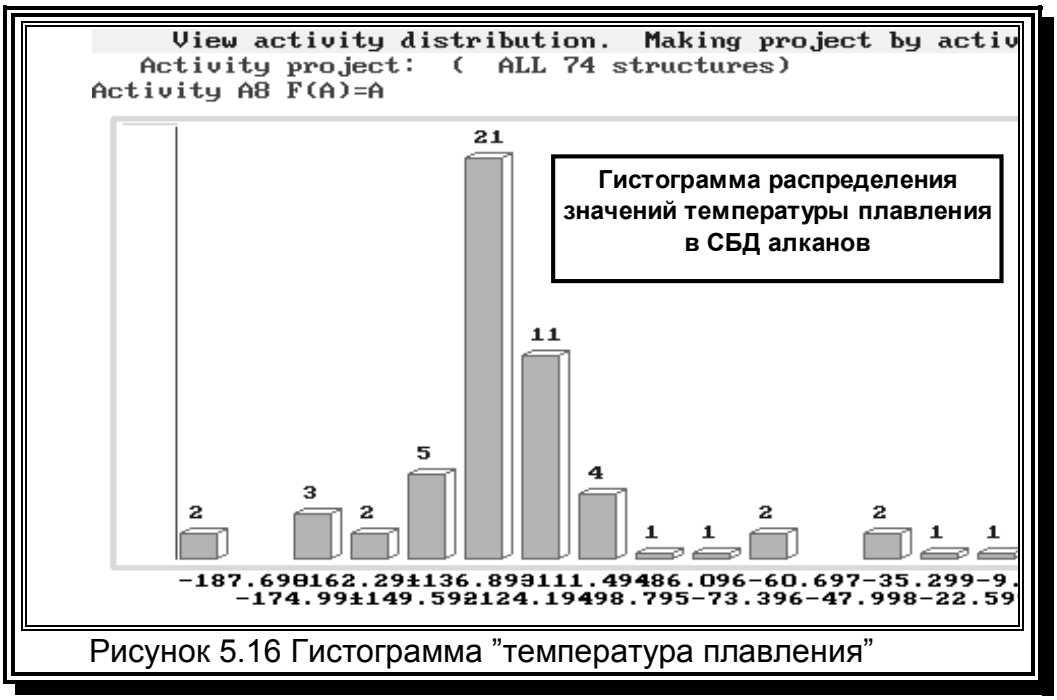


Рисунок 5.16 Гистограмма ”температура плавления”

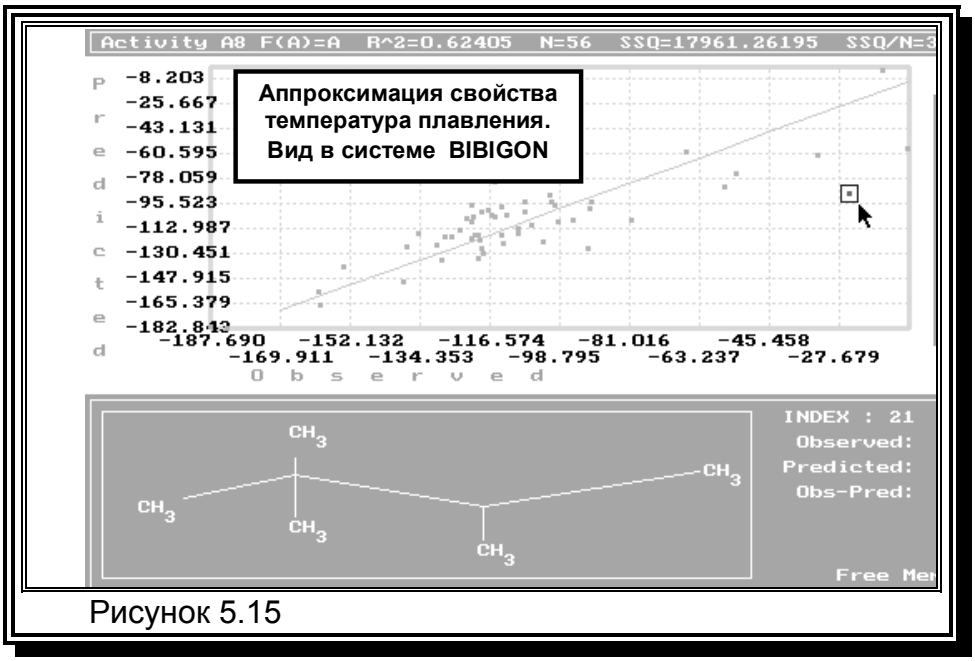


Рисунок 5.15

Скользящий Контроль: **N = 56, K = 7, R^2_{CV} = 0.574, S = 23.67, F = 1.21**

1. **4586.86 (183.43) CONST**
2. **87.344 (21.82) (W: 1600) Chain F(X)=inf.X C 4s*C 4s*C 4s*C 4s*H 1s***
3. **-2193.8 (105.30) (W: 1320) Chain F(X)=av.X C 4s*C 4s*H 1s***
4. **6020.0 (265.03) (W: 1320) Chain F(X)=inf.X C 4s*C 4s*H 1s***
5. **-2121.4 (207.69) (W: 956) Chain F(X)=av.X H 1s***
6. **-30.1622 (9.065) (W: 2137) Chain F(X)=av.X H 1s*C 4s*C 4s*C 4s*H 1s***
7. **-0.0039 (0.0011) (W: 2137) Chain F(X)= X¤ H 1s*C 4s*C 4s*C 4s*H 1s***

Таким образом, прогностичное качество линейной модели аппроксимации температуры плавления алканов низкое, несмотря на применение нелинейных преобразований дескрипторов и использования топологических индексов.

Приведенные выше результаты скользящего контроля **QSAR**-моделей, построенных на цепочках маркированных атомов разной длины, свидетельствуют о высокой прогностичностиэтих моделей для всех свойств, кроме температуры плавления. Это свойство известно, как существенно нелинейное - небольшие изменения структурной формулы могут приводить к резкому изменению величины свойства.

Далее применялась следующая методика расчетов, направленная на поиск оптимальной для данного свойства маркирования атомов и максимальной длины цепочки. Для каждого свойства проводилось четыре группы расчетов:

- с отключенной маркировкой (маркировка обозначена как ***),
- с включенным p- маркером (обозначение - **X**),**
- с включенными p- и b- маркерами (обозначение - **XX*),** и
- со всеми включенными маркерами (обозначение - **XXX**).

Для каждого типа маркировки проводились рассчеты с варьированием максимальной длиной атомных цепочек от **1** до **4**. Критерием остановки селекций на каждом этапе рассчета служило максимальное количество предикторов в модели (оно задавалось как (**N/5**), где **N** - число соединений в СБД) и минимальный порог изменения $\mathbf{R^2}$ от **i**-ой селекции к **(i+1**)-й (**0.005**). После завершения процесса селекция проводилась процедура скользящего контроля для всех уравнений последней селекции. Результаты скользящего контроля в зависимости от изменения маркировки и максимальной длины атомных цепочек приведены на графиках для каждого рассмотренного свойства. По значению $\mathbf{R^2}$, полученному при скользящем контроле отбиралась лучшая кодировка и максимальная длина цепочки. Из нескольких вариантов лучших кодировок и длин, имеющих одинаковый $\mathbf{R^2}$, выбирался самый простой (минимальная длина и простейшая кодировка), что в принципе обеспечивает лучшую прогностичность модели (более широкую применимость).

С использованием лучшей для данного свойства кодировкой и максимальной длиной цепочки проводились расчеты, призванные продемонстрировать, что приведенная модель действительно является лучшей по критерию $\mathbf{R^2_{cv}}$ (cross-validation). Для этого проводилось построение моделей с числом дескрипторов от 1 до заданного числа. После чего строились графики зависимости $\mathbf{R^2}$ и $\mathbf{R^2_{cv}}$ от числа дескрипторов в модели **(K).**

5.2 Энтальпия образования веществ разных химических классов.

Обучающая выборка введена в ЭВМ А.Н.Зефировым на основе статьи [338]. Выборка содержала **525** соединений. Проводился расчет энтальпии образования веществ.

Рисунок 5.17 Гистограмма “энтальпия образования

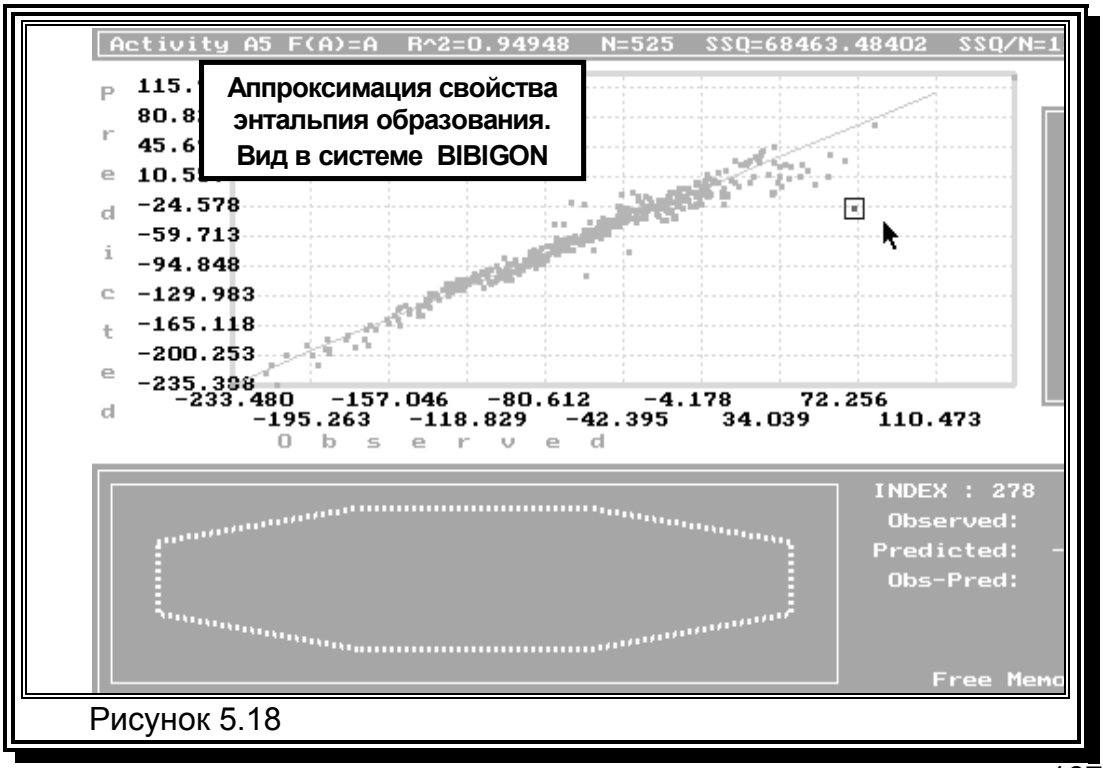


Рисунок 5.18

"Скользящий контроль": **N=525, K=11, R^2_{CV} =0.948, S=11.95, F=858.4**

Коэффициенты лучшей **QSAR**-модели (маркировка **XXX**, длина цепочек **<=4**):

1. 21.123 (0.206) CONST
2. 21.290 (0.152) (W: 61) Chain F(X)= X C 1dc
3. 11.624 (0.095) (W: 90) Chain F(X)= X C 2dc
4. 13.473 (0.079) (W: 74) Chain F(X)= X C 2ds
5. -8.591 (0.106) (W: 281) Chain F(X)= X C 3ss
6. 5.751 (0.078) (W: 121) Chain F(X)= X C 3ssC 3ssC 3ss
7. -33.831 (0.089) (W: 105) Chain F(X)= X O 1dc
8. -36.173 (0.077) (W: 101) Chain F(X)= X O 1sc
9. -29.022 (0.068) (W: 98) Chain F(X)= X O 2sc
10. -28.778 (0.084) (W: 68) Chain F(X)= X O 2ss
11. -26.867 (0.136) (W: 4597) Chain F(X)= X _LEN1
12. 21.015 (0.132) (W: 4405) Chain F(X)= X _LEN2

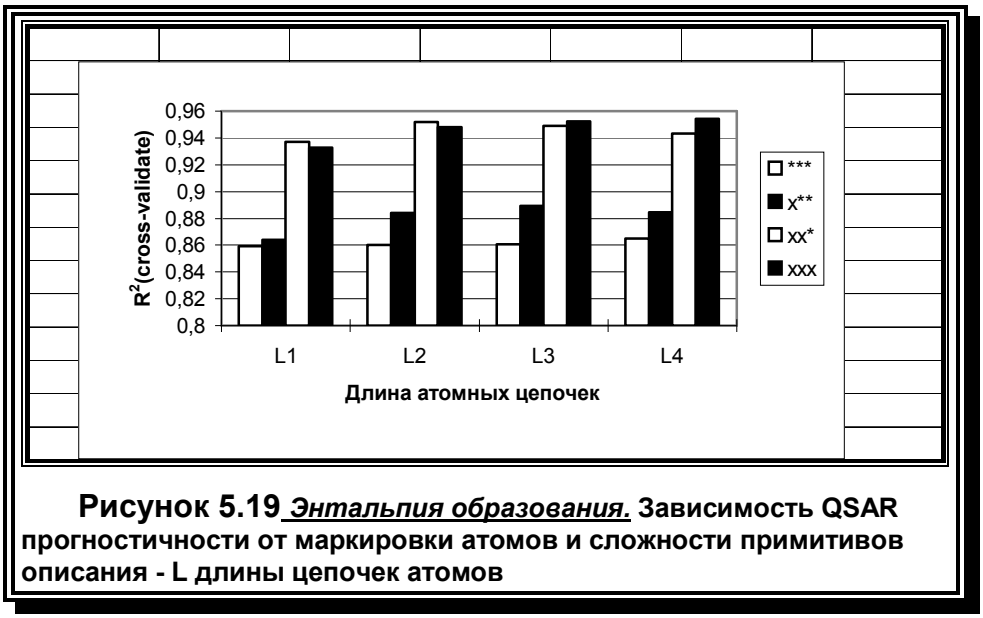


Рисунок 5.19 ***Энтальпия образования.*** **Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов**

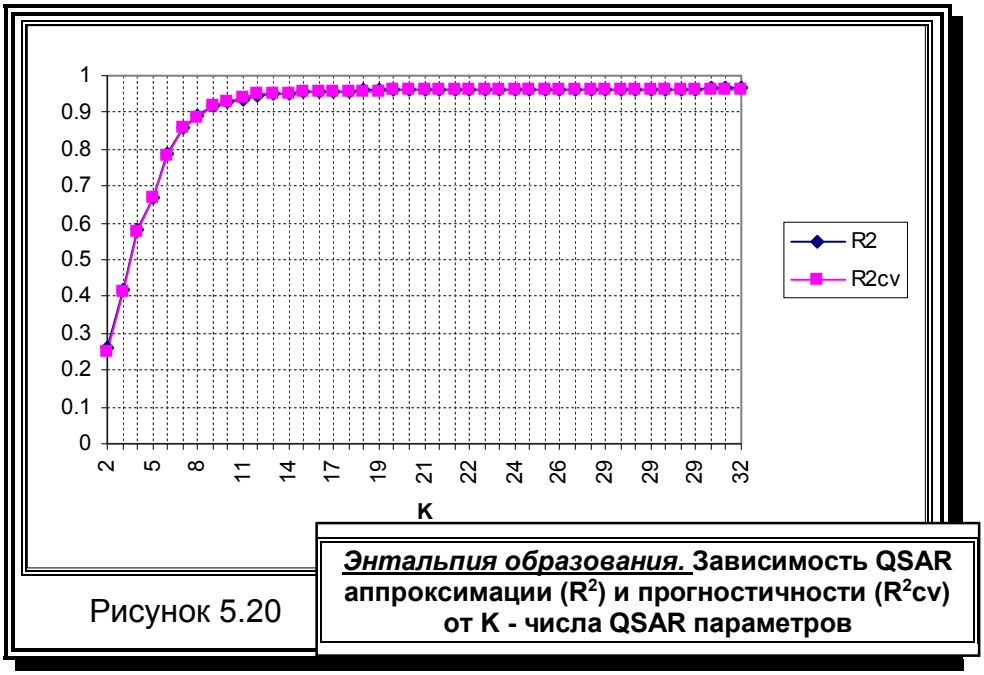


Рисунок 5.20

Энтальпия образования. **Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров**

5.3 Поляризуемость молекул.

Обучающая СБД была введена на Химфаке МГУ А.Н.Зефировым и Н.С.Зефировым на основе работы[339]. Обучающая выборка содержала 293 соединения, из них 279 были органическими, а 14 - неорганическими веществами (не содержали углерод). Гистограмма свойства “поляризуемость молекул” имеет следующий вид:

Рисунок 5.21 Гистограмма “поляризуемость молекул”

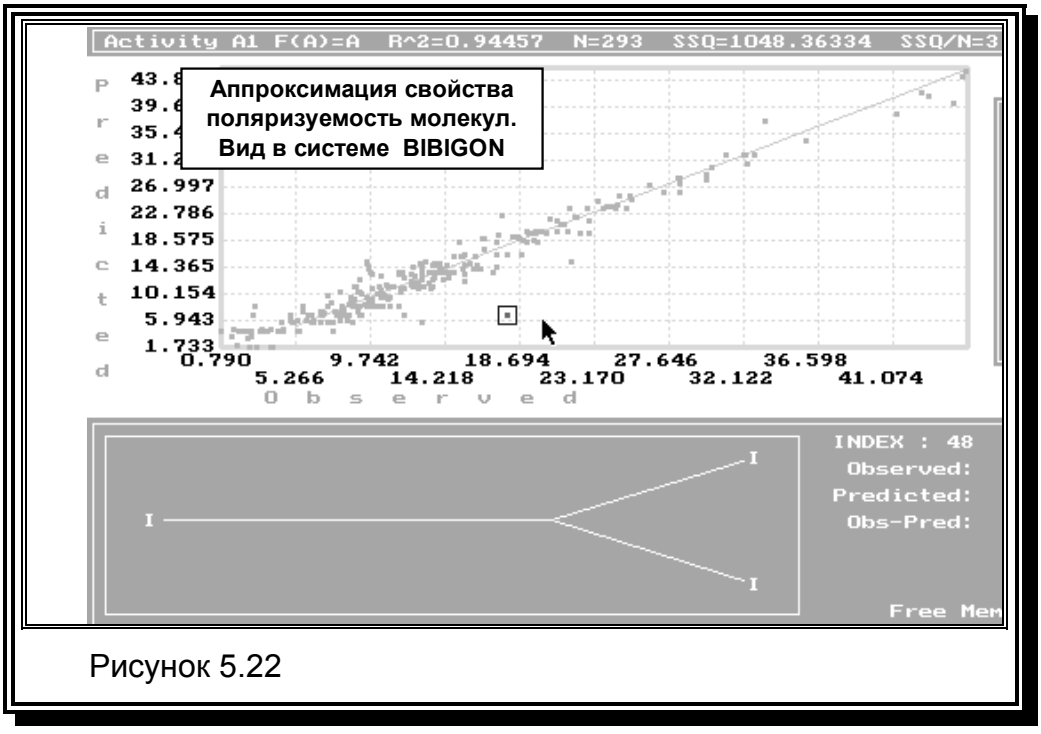


Рисунок 5.22

"Ñêîëüçyùèé êîíòðîëü": **N=293, K=4, R^2_{CV} =0.972, S=1.37, F=2575.63**
Ëó÷øàÿ **QSAR**-ìîäåëü (êîäèðîâêà XX*, äëèíà öåïî÷åê = 1):

1. **0.6696 (0.024) CONST**
2. **-1.7991 (0.008) F(X)= X F 1s***
3. **3.3434 (0.027) F(X)= X I 1s***
4. **-1.7945 (0.017) F(X)= X O 1d***
5. **1.7026 (0.002) F(X)= X _LEN1**

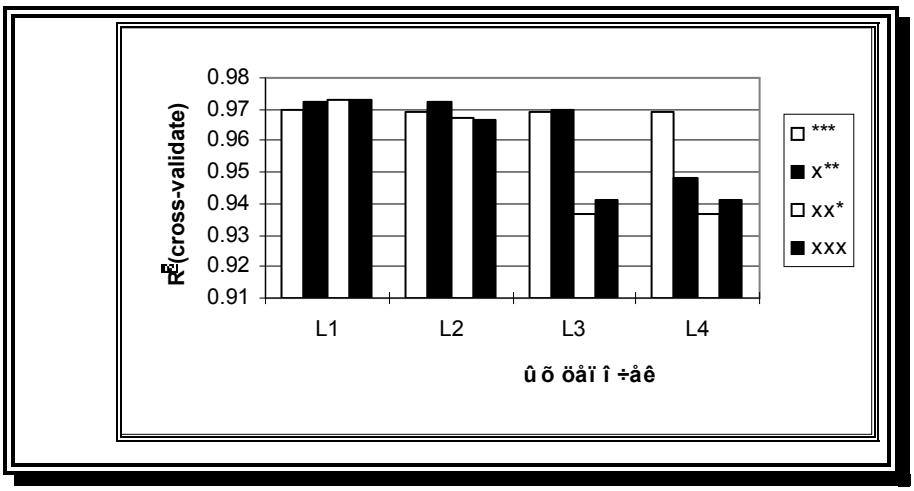


Рисунок 5.23 Поляризуемость молекул. Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

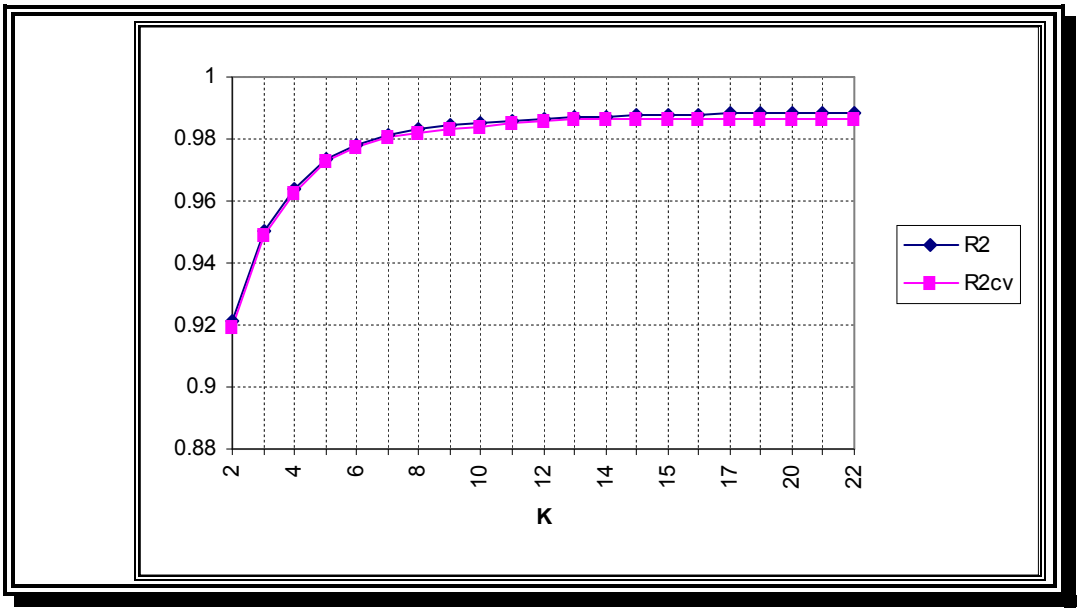


Рисунок 5.24 Поляризуемость молекул. Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

5.4 Температура кипения для фуранов/тетрафуранов и тиофенов.

Обучающая СБД была введена на Химфаке МГУ А.Н.Зефировым и Н.С.Зефировым на основе работы[340]. Выборка содержала **323** соединения, из них фураны - **88** соединений, тетрагидрофураны и лактоны - **120** соединений, тиофены и тетрагидротиофены - **123** соединения. Гистограмма свойства “температура кипения” имеет следующий вид:

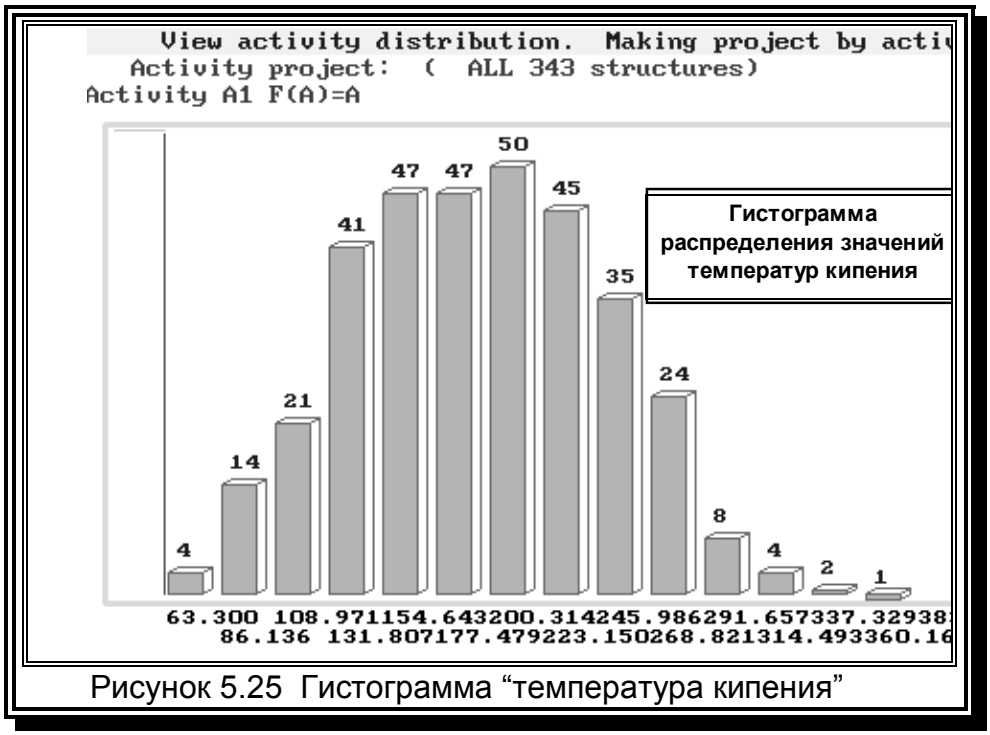


Рисунок 5.25 Гистограмма “температура кипения”

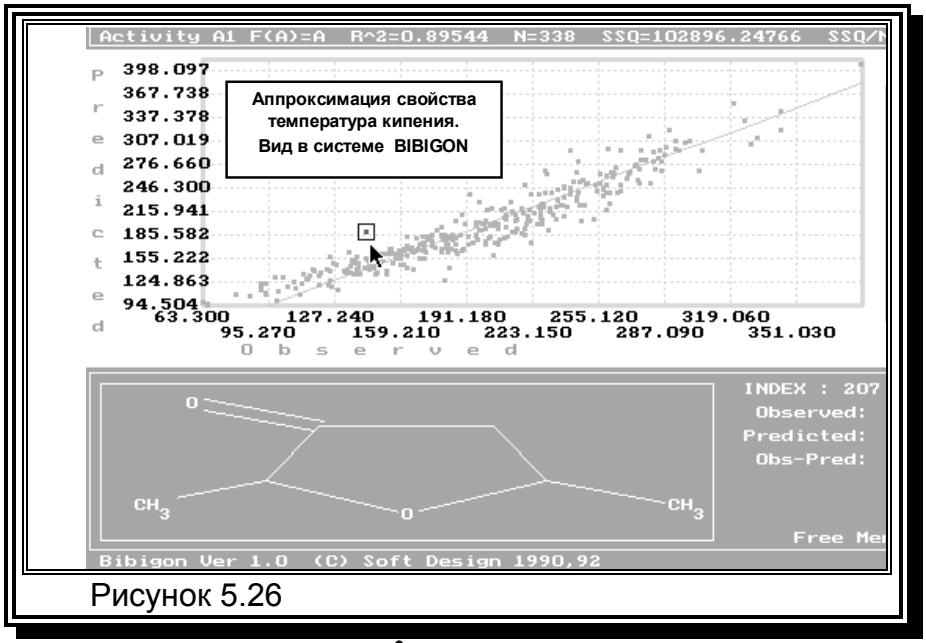


Рисунок 5.26

"Ñêîëüçÿùèé êîíòðîëü": **N=343, K=10, R^2_{CV} =0.897, S=18.38, F=293.12.**

Ëó÷øàÿ **QSAR**-ìîäåëè (êîäèðîâêà **XXX**, äëèíà öåïî÷åê **<=2**):

1. 37.631 (0.53) CONST
2. 30.649 (0.16) F(X)= X BR1sc
3. 3.720 (0.07) F(X)= X C 2ar
4. -25.749 (0.43) F(X)= X C 3dcO 2sc
5. -38.970 (0.67) F(X)= X C 3dsC 4ss
6. 28.147 (0.45) F(X)= X C 3dsO 1dc
7. 36.499 (0.28) F(X)= X O 1dc
8. 42.346 (0.69) F(X)= X O 1sc
9. -32.057 (0.25) F(X)= X O 2ss
10. 22.111 (0.11) F(X)= X _LEN2
11. -5.245 (0.07) F(X)= X _LEN3

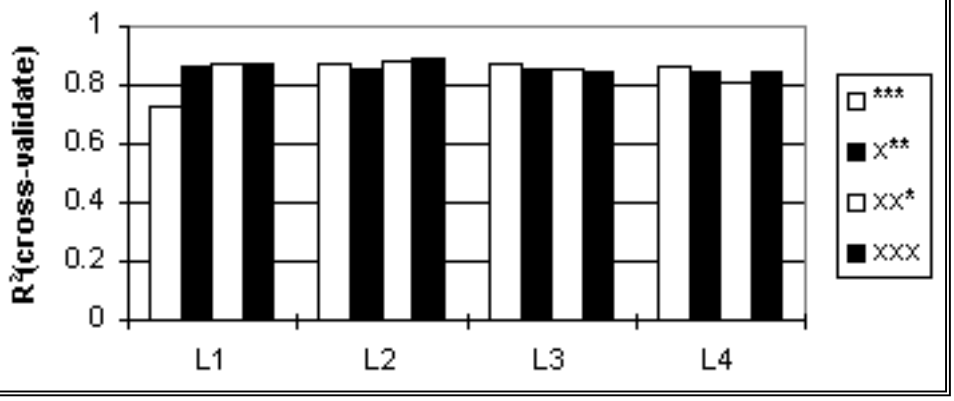


Рисунок 5.28 Температура кипения. Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

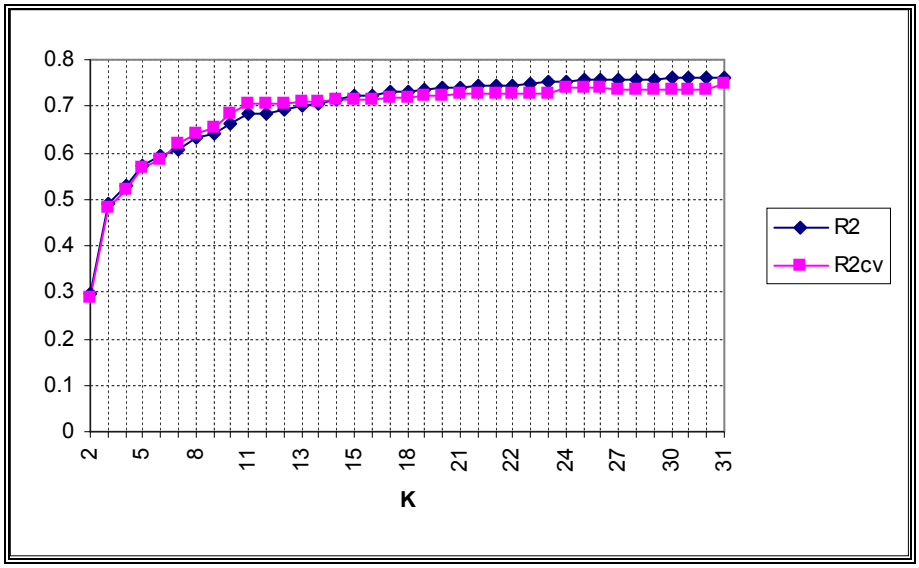


Рисунок 5.27 Температура кипения. Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

5.5 Липофильность замещенных бензолов.

Îáó÷àþùàÿ ÑÁÄ áûëà ââåäåíà â ÝÂÌ Ïîíîìàðåâîé Ë.À. íà îñíîâå ðàáîòû [341]. Îáó÷àþùàÿ âûáîðêà ñîäåðæàëà 172 ñîåäèíåíèé.

"Ñêîëüçÿùèé êîíòðîëü": **N=172, K=9, R^2_{cv}=0.926, S=0.334, F=229.56**

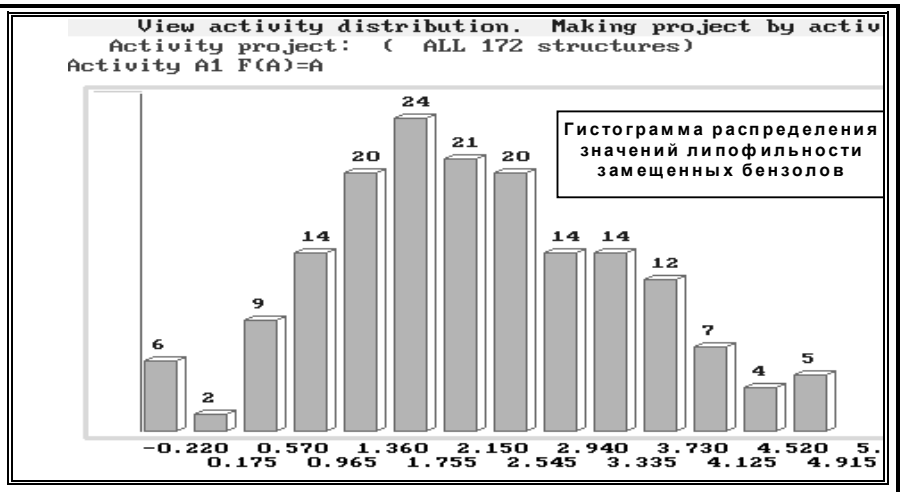


Рисунок 5.29 Гистограмма “липофильность замещенных бензолов”

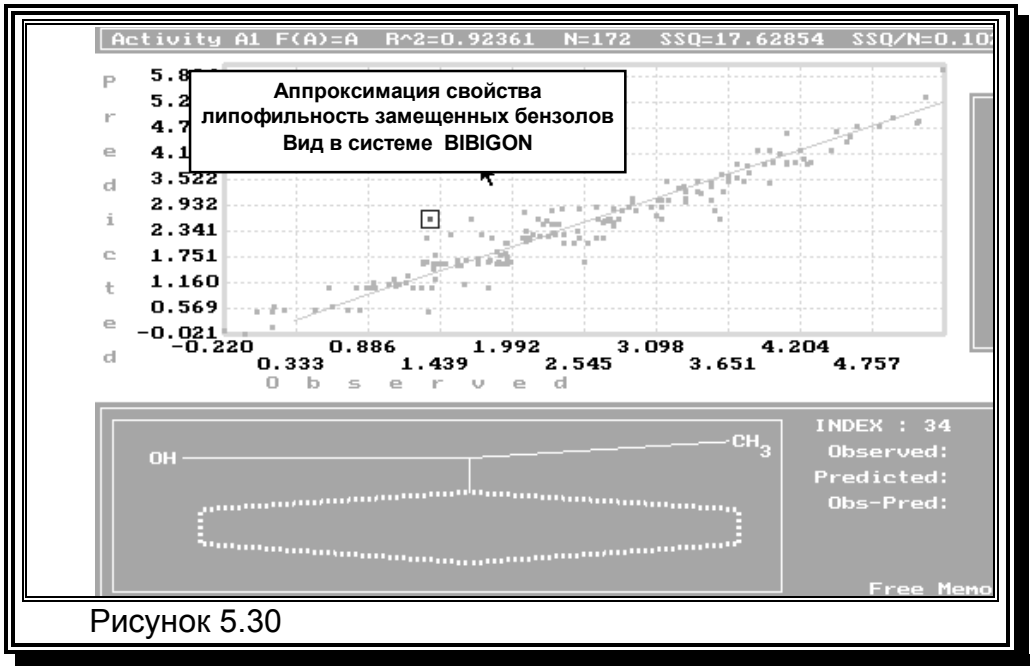
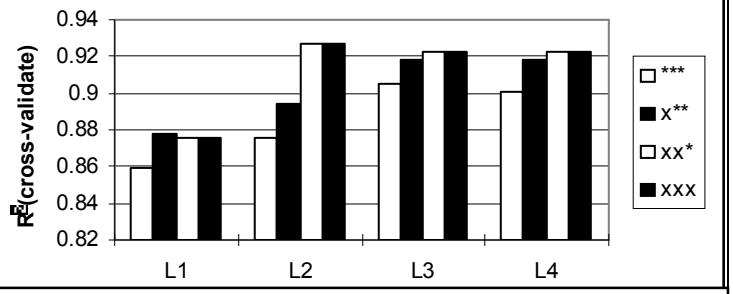


Рисунок 5.30

Ëó÷øàÿ **QSAR**-ìîäåëü (êîäèðîâêà **XX***, äëèíà öåïî÷êè **<=2**):

1. -0.9201 (0.046) CONST
2. 0.9775 (0.007) F(X)= X C 1s*
3. 0.8552 (0.016) F(X)= X C 1s*C 2s*
4. -0.7339 (0.017) F(X)= X C 1s*C 3d*
5. 0.5039 (0.009) F(X)= X C 2a*
6. -0.5416 (0.011) F(X)= X C 2s*
7. 1.1488 (0.018) F(X)= X C 2s*C 2s*
8. 1.1931 (0.008) F(X)= X Cl1s*
9. -0.4362 (0.009) F(X)= X N 1s*
10. -0.4913 (0.008) F(X)= X O 2s*



Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

Рисунок 5.31 Липофильность замещенных бензолов

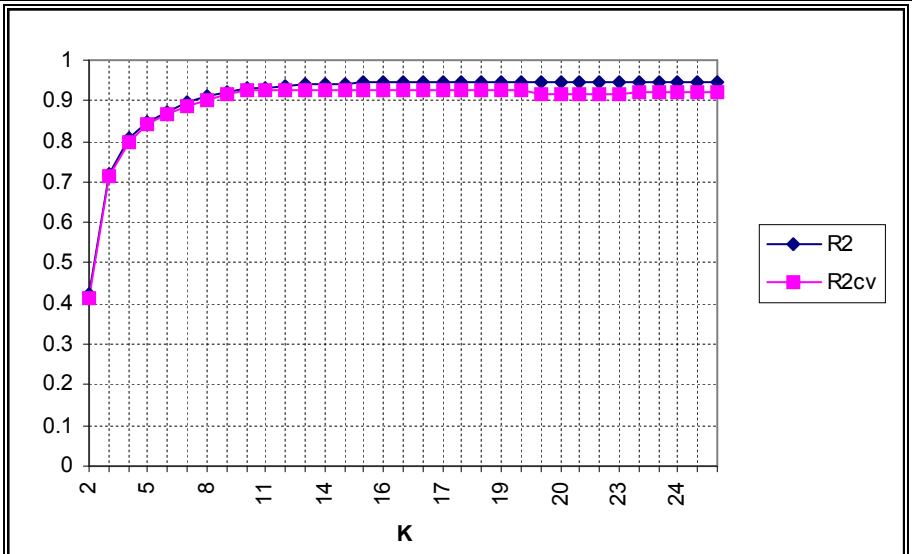


Рисунок 5.32 Липофильность замещенных бензолов. Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

На это выборке нами было исследовано влияние различных маркировок атомов на качество получаемой аппроксимации свойства. Результаты объединены в Таблице 5.2

Таблица 5.2 Сравнение различных маркировок атомов в QSAR моделях аппроксимации липофильности замещенных бензолов. Приведены результаты регрессионных моделей. Лучшая маркировка атомов выделена.

маркировка ***	K = 17	R^2= 0.945	S=0.299	F = 151.77
маркировка X**	K = 17	R^2= 0.970	S=0.219	F = 291.13
** маркировка *X**	*K = 16*	*R^2= 0.976*	*S=0.194*	*F = 393.53*
маркировка **X	K = 17	R^2= 0.969	S=0.223	F = 278.67
маркировка XX*	K = 17	R^2= 0.973	S=0.210	F = 316.06
маркировка X*X	K = 17	R^2= 0.967	S=0.231	F = 259.15
маркировка *XX	K = 17	R^2= 0.970	S=0.218	F = 292.44
маркировка XXX	K = 18	R^2= 0.946	S=0.298	F = 144.33

5.6 Чувствительность к удару нитросоединений.

Îáó÷àþùàÿ âûáîðêà, ïðåäñòàâëåíiàÿ ñîòðóäíèêàìè ëàáîðàòîðèè íèòðîñîåäèíåíèé ÈÎÕ ÐÀÍ, ñîäåðæàëà **155** íèòðîñîåäèíåíèé ðàçëè÷íûõ õèì÷åñêèõ êëàññîâ. Äëÿ ïðèâåäåíèÿ ðàñïðåäåëåíèÿ àêòèâíîñòè ê íîðìàëüíîìó äëÿ äàííîé âûáîðêè âû÷èñëåíiÿ ïðîãíîçèðîâü äëÿ ïðåíðàçîâàííîé àêòèâíîñòè **Ln(H_{50}).**

"Ñêîëüçÿùèé êîíòðîëü": **N=155, K=18, R^2_{CV} =0.689, F=16.9, S=0.667**.
Êîýôôèöèåíòû ëó÷øåé **QSAR**-ìîäåëè (êîäèðîâêà **X**,** äëèíà öåïî÷åê **<= 4**):

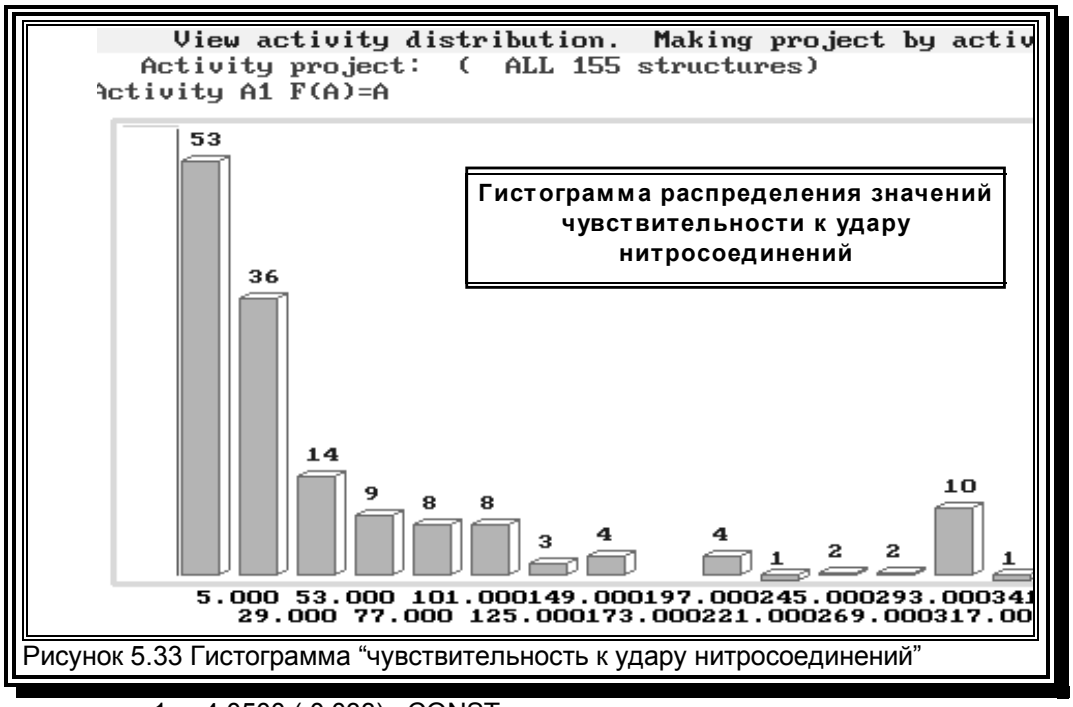


Рисунок 5.33 Гистограмма "чувствительность к удару нитросоединений"

1. 4.0500 (0.033) CONST

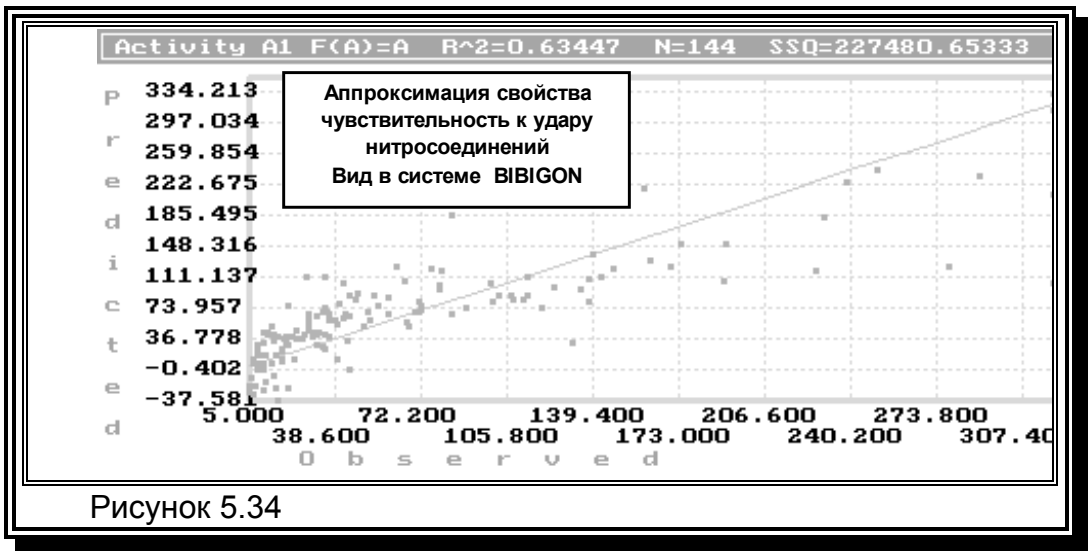
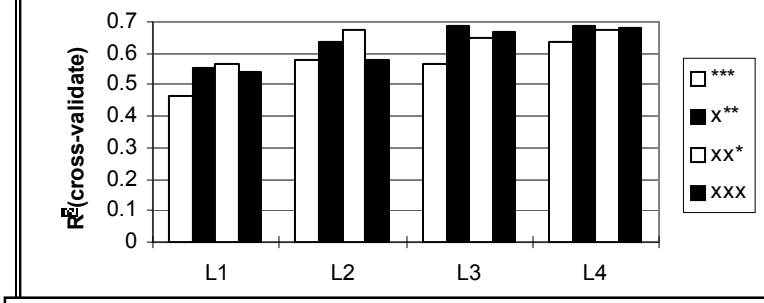


Рисунок 5.34

2. 0.2881 (0.012) F(X)= X C 1**
3. 0.4239 (0.020) F(X)= X C 1**C 3**C 3**C 2**
4. 0.6244 (0.023) F(X)= X C 1**C 4**C 2**O 2**
5. 0.9032 (0.146) F(X)= X C 1**N 3**O 1**
6. 0.3313 (0.022) F(X)= X C 1**O 2**C 3**C 3**
7. 0.2905 (0.007) F(X)= X C 2**

8. -0.4638 (0.030) F(X)= X C 2**C 2**O 2**N 3**
9. 0.2312 (0.023) F(X)= X C 2**C 3**C 3**O 1**
10. 0.2042 (0.012) F(X)= X C 2**C 4**
11. 0.8630 (0.034) F(X)= X C 2**N 2**N 2**
12. 0.9130 (0.032) F(X)= X C 2**N 2**N 3**C 3**
13. -2.0912 (0.148) F(X)= X C 2**N 3**N 2**N 2**
14. 0.1415 (0.005) F(X)= X C 3**
15. 0.6939 (0.017) F(X)= X N 1**
16. 1.7171 (0.278) F(X)= X N 2**N 2**C 3**O 1**
17. -0.3881 (0.007) F(X)= X N 3**
18. 0.2806 (0.014) F(X)= X N 3**N 3**
19. -0.2222 (0.004) F(X)= X O 1**N 3**O 2**



Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

Рисунок 5.35 Чувствительность к удару нитросоединений

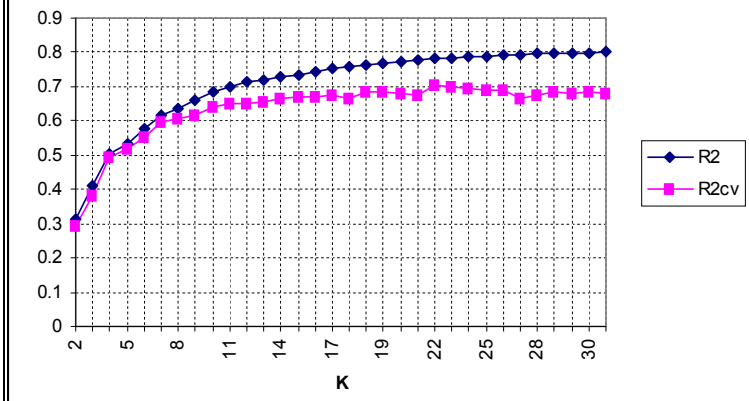


Рисунок 5.36 Чувствительность к удару нитросоединений. Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

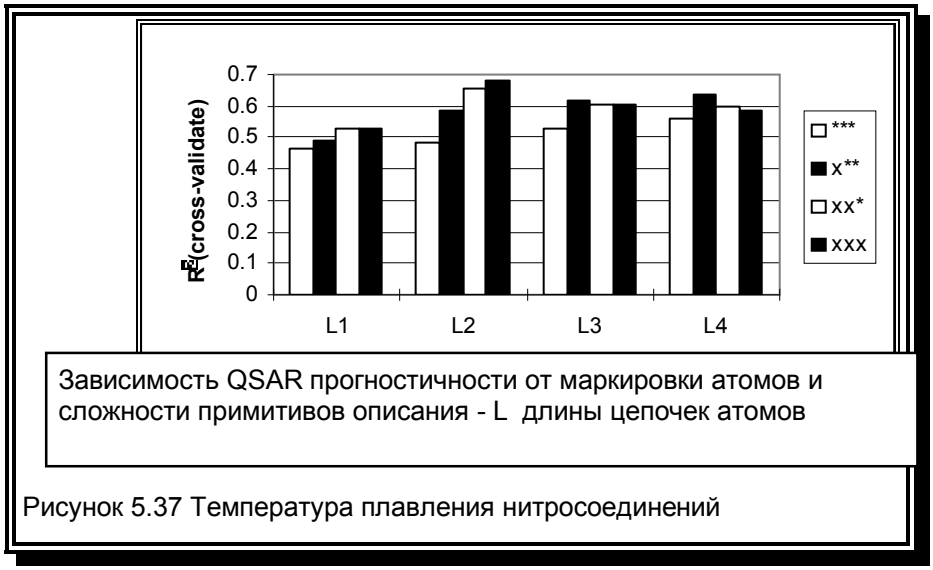
5.7 Температура плавления нитросоединений.

Âûáîðêà, ïðåäñòàâëåíèàÿ ñîòðóäíè÷àíèå ëàáîðàòîðèè íèòðîñîåäèíåíèé ÈÎÕ ÐÀÍ, ñîäåðæàëà **225** íèòðîñîåäèíåíèé.

"Ñêîëüçÿùèé êîíòðîëü": **N=225, K=20, R^2_{CV} =0.679, S=48.35, F=21.71**

Êîýôôèöèåíòû ëó÷øåé **QSAR**-ìîäåëè (êîäèðîâêà XXX, äëèíà öåïî÷åê <=2):

1. 23.698 (1.03) CONST
2. -60.612 (3.27) F(X)= X BR1sc
3. -32.646 (1.84) F(X)= X C 1scC 2sc
4. -34.756 (1.41) F(X)= X C 1scC 3sc
5. -6.746 (0.23) F(X)= X C 2scC 2sc
6. 45.000 (0.91) F(X)= X C 2scN 2sc
7. 43.396 (3.22) F(X)= X C 3asN 1sc
8. -48.935 (1.30) F(X)= X C 3asN 3sc
9. 93.016 (0.99) F(X)= X C 3dcN 1sc
10. -55.999 (0.78)F(X)= X C 3scN 3sc
11. -32.058 (2.13)F(X)= X C 3ssN 3wc
12. 70.791 (4.85)F(X)= X C 4scC 4sc
13. -16.021 (0.26)F(X)= X C 4scN 3wc
14. -155.591(20.80F(X)= X C 4ssN 3wc
15. -61.847 (1.67) F(X)= X CL1sc
16. 31.732 (1.42) F(X)= X N 2arN 2ar
17. -75.233 (5.30) F(X)= X N 2dcN 2dc
18. 14.204 (0.86) F(X)= X N 3sc
19. 32.293 (0.33) F(X)= X N 3ssN 3wc
20. -18.658 (0.21) F(X)= X O 2sc
21. 3.849 (0.03) F(X)= X _LEN4



Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

Рисунок 5.37 Температура плавления нитросоединений

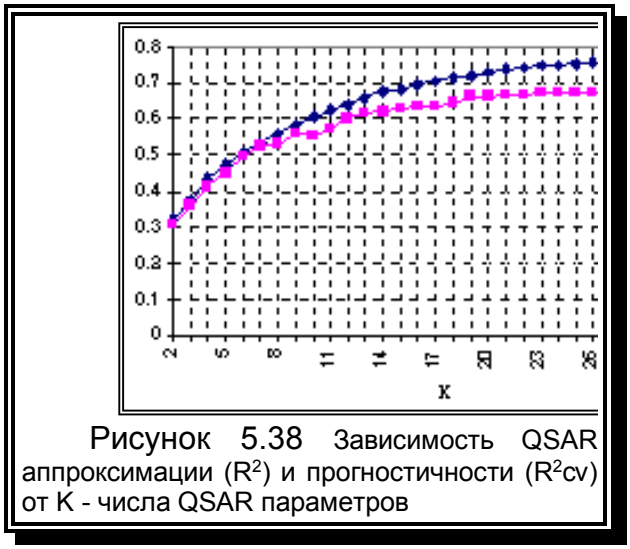


Рисунок 5.38 Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

5.8 Хроматографическая подвижность антрациклиновых антибиотиков ряда даунорубицина.

Âûáîðêà âêëþ÷àåò â ÝÂÌ Ïîíèìàðåâîé Ë.À. íà îñíîâå äàííûõ ÂÍÈÈ Íîâûõ Àíòèáèîòèêîâ ÐÀÌÍ[342]. Îáó÷àþùàÿ ÑÁÄ ñîäåðæàëà **87** ñîåäèíåíèé. Ìîëåêóëû àíòèáèîòèêîâ ïðåäñòàâëÿþò ñîáîé äîñòàòî÷íî ñëîæíûå ñòðóêòóðû, èìåþùèå ïîñòîÿííóþ ÷àñòü, ñîñòîÿùóþ èç êîëüöåâîé ñèñòåìû ñ ÷åòûðüìÿ öèêëàìè. Â ñâÿçè ñ ýòèì êëàññè÷åñêèé ðàñ÷åò ëèíîéíîé çàâèñèìîñòè òåùåñòâ çàòðóäíåí, à ïîðîé è íåâîçìîæåí èç-çà íåîáõîäèìîñòè ìíîæåñòâåííîãî ó÷åòà èìåþùèõñÿ ôðàãìåíòîâ è èõ âçàèìîâëèÿíèÿ äðóã íà äðóãà. Ëèíîéíûíîñòü ýòèõ âåùåñòâ õîðîøî êîððåëèðóåò ñ èõ õðîìàòîãðàôè÷åñêîé ïîâåäåíèíîñòüþ [343,344].

"Ñêîëüçÿùèé êîíòðîëü": **N=87, K=16, R^2_{CV} =0.936, S=0.058, F=64.79**

Êîýôôèöèåíòû ëó÷øåé **QSAR**-ìîäåëè (êîäèðîâêà **XXX**, äëèíà öåïî÷åê **<= 3**):

1. -0.09515 (0.0194) CONST
2. 0.10244 (0.0032) F(X)= X C 1scC 3dcO 1dc

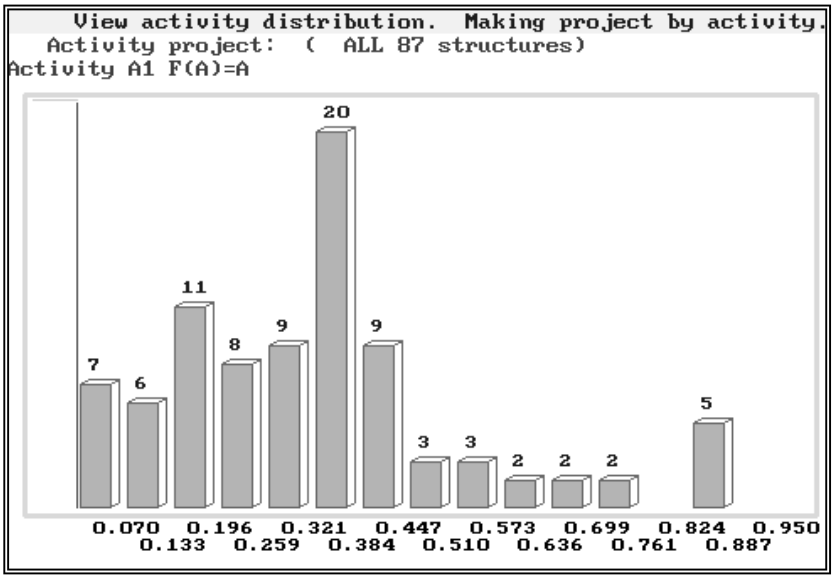


Рисунок 5.39 Гистограмма свойства “хроматографическая подвижность”

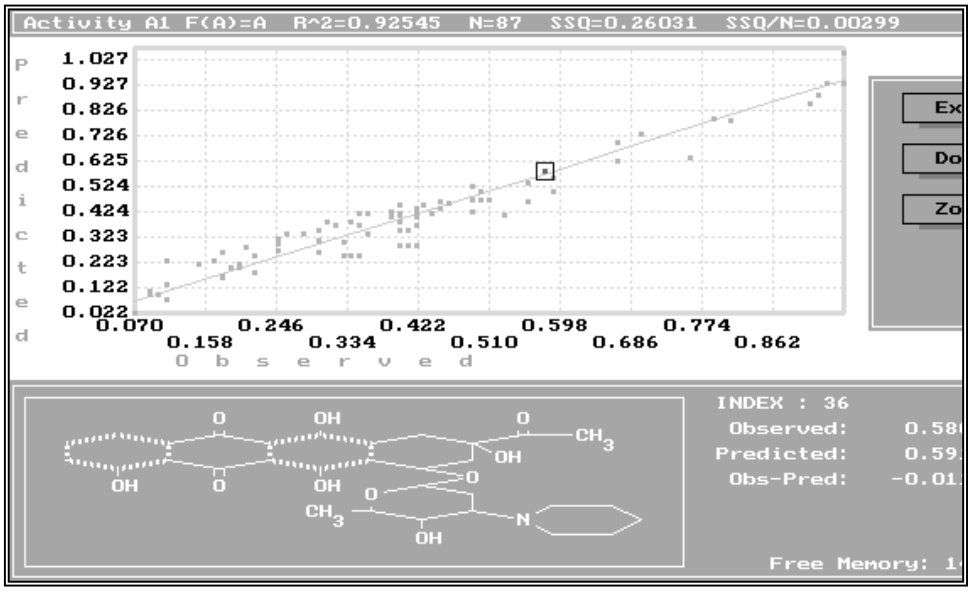
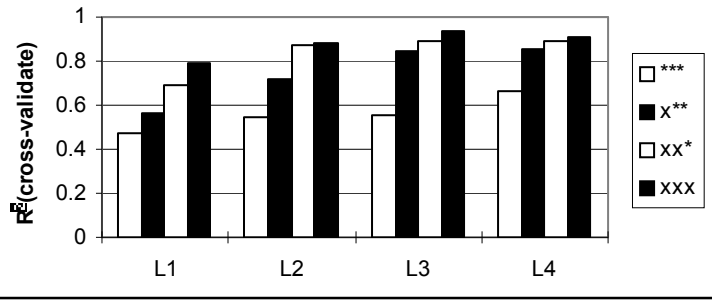


Рисунок 5.40

3. -0.12667 (0.0074) F(X)= X C 1scC 3scC 3dc
4. 0.03460 (0.0019) F(X)= X C 1scC 4scO 2sc
5. -0.15394 (0.0033) F(X)= X C 1scN 3sc
6. -0.10090 (0.0087) F(X)= X C 2dcO 2sc
7. 0.03170 (0.0029) F(X)= X C 2scC 2sc
8. 0.10683 (0.0074) F(X)= X C 2scCL1sc
9. 0.02607 (0.0009) F(X)= X C 2ss
10. 0.13975 (0.0032) F(X)= X C 2ssC 2ssO 2ss
11. 0.07789 (0.0015) F(X)= X C 3as
12. -0.11099 (0.0033) F(X)= X N 1sc
13. -0.17616 (0.0036) F(X)= X N 2arC 2arN 3as
14. 0.39010 (0.0058) F(X)= X N 2scC 3dcN 2sc
15. -0.07167 (0.0033) F(X)= X N 2scC 3dcO 1dc
16. -0.17347 (0.0105) F(X)= X N 2ss
17. -0.06381 (0.0020) F(X)= X O 1sc



Зависимость QSAR прогностичности от маркировки атомов и сложности примитивов описания - L длины цепочек атомов

Рисунок 5.42 Хроматографическая подвижность

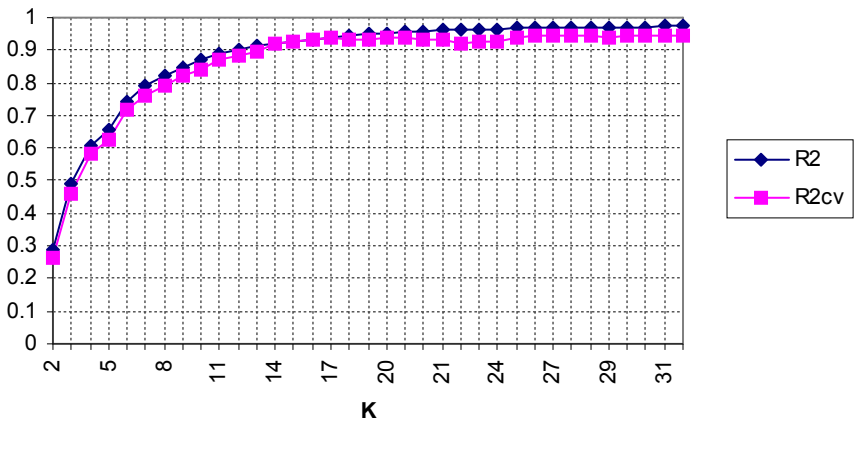


Рисунок 5.41 Хроматографическая подвижность. Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от K - числа QSAR параметров

5.9 Анализ большой СБД по липофильности.

Известно, что величина **LogP** - логарифм показателя распределения вода/октанол - хорошо коррелирует со многими биологическими свойствами веществ [345]. Аспирантами Химфаыка МГУ, Ощепковым Э.Н. и Колововановым Э.А.- в течение трех лет была создана база данных, содержащая более **3500** химических структур, с охарактеризованными значениями **LogP**. СБД по **LogP** была создана на основе **180** литературных источников. Многие соединений были описаны одновременно в нескольких источниках и имели явные противоречия в данных.

При создании **QSAR**-модели для оценки **LogP** был использован постулат о том, что величина **LogP** является аддитивно-конституитивной. Это означает, что величину **LogP** для какого-либо соединения можно представить в виде суммы, которая содержит:

1.Аддитивные вклады структурных фрагментов (атомов и/или групп атомов) молекулы.

2.Поправочные члены, отражающие взаимное влияние определенных фрагментов молекулы на вносимый ими вклад в общую величину **LogP** (учет внутримолекулярных взаимодействий).

Разбиение М-графа на фрагменты производится на основании:

Предопределенной внутренней библиотеки, содержащей более 500 структурных фрагментов с определенными для них аддитивными инкрементами.

Списка правил для объединения нескольких смежных фрагментов в один новый. Аддитивный инкремент для такого “нового” фрагмента рассчитывается на основании дополнительного **QSAR**-уравнения.

Внутримолекулярные взаимодействия, рассматриваемые в **QSAR**-модели, учитывают взаимное влияние структурных фрагментов на их аддитивные инкременты. На основании того, через какие углеродные атомы происходит взаимодействие, внутримолекулярные взаимодействия разделены на четыре типа:

- алифатические - учитываются между группами, соединенными углеродной цепочкой длиной до 4 атомов, в которой имеется хотя бы один sp3 углеродный атом.

- ароматические - учитываются между группами, соединенными углеродной цепочкой любой длины, в которой все атомы цепочки принадлежат одной и той же ароматической системе.
- винильные - учитываются между геминальными и вицинальными заместителями двойной связи [346].
- кольцевые - взаимодействия между ароматическими кольцами через мостиковые связи.

Значения поправок на внутримолекулярные взаимодействия находятся во внутренней библиотеке, полученной на основе обработки экспериментальных данных и содержащей около **2,000** цепочек, либо рассчитываются для “новых” взаимодействий по специальному **QSAR**-уравнению.

В **QSAR**-модель включен алгоритм расчета 95% доверительного интервала для оценки **LogP**. Значение доверительного интервала рассчитывается исходя из ошибок определения каждого из составляющих **LogP** инкрементов. В свою очередь, значения ошибок определения инкрементов или определены заранее (для фрагментов из библиотеки), или рассчитываются по специальному **QSAR**-уравнению для “новых” фрагментов. Экспериментальная ошибка определения **LogP** обычно составляет **± 0.1 - ± 0.3** единиц шкалы **LogP**.

В модель были также включены правила для фильтрации М-графов. Расчет **LogP** не производится для следующих соединений:

1.Соединений, включающих химические элементы, отличных от: **C, H, D, O, S, P, N, F, Cl, Br, I, Se, Si, Ge, Pb, Sn, As, B.**

2.Соединений, содержащих хотя бы один химический элемент с валентностью и/или координационным числом, в которых этот элемент не встречается во внутренней библиотеке фрагментов

3.Ионных форм соединений.

Выборочные результаты корреляций экспериментальных значений **LogP** от расчитанных при помощи построенной **QSAR**-модели приведены ниже. Расчеты проведены совместно с аспирантами Митюшевым Д.Ф., Ощепковым Э.Н. и Колововановым Э.А.

1) При использовании всех соединений СБД получен результат:

$$LogP_{exp} = 0.0751\ (\pm 0.0090) + 0.9482\ (\pm 0.0035)\ LogP_{calc}$$

N=3695, R=0.9759, S=0.3658

2) При исключении из расчетов “выпадающих точек” получен результат:

$$LogP_{exp} = 0.0290\ (\pm 0.0052) + 0.9888\ (\pm 0.0021)\ LogP_{calc}$$

N=3601, R=0.9922, S=0.2088, где **N** - количество точек, **R** - коэффициент корреляции, **S** - стандартное отклонение

Таким образом, **QSAR**-модель дает для 90% соединений из СБД отклонение предсказанного значения от экспериментального не более 0.1, что практически является ошибкой эксперимента измерения **LogP** [347]. Тестирование **QSAR**-модели показало сохранение прогностичных характеристик оценок на уровне экспериментальной ошибки измерения **LogP**.

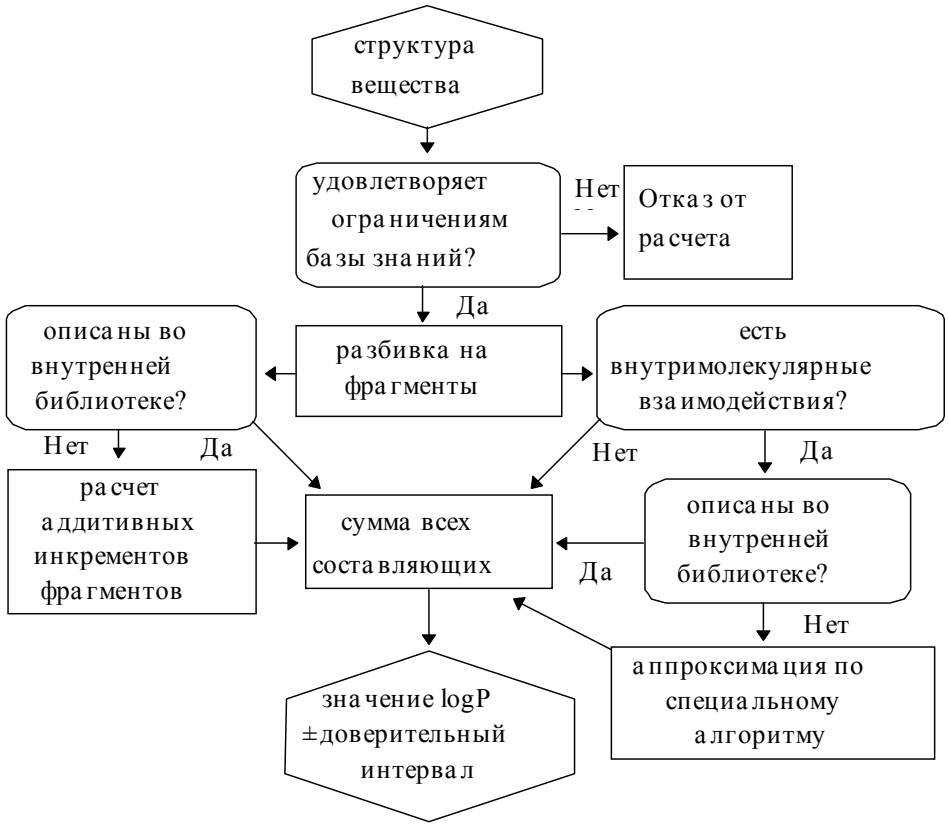


Рисунок 5.43 Схема работы механизма вывода оценок LogP, исходя из структуры соединения, являющейся запросом к базе знаний "LogP". Этап разбивки на фрагменты включает "кодирование фрагментов", которое задает способ маркирования атомов. Изменение условий маркирования существенно влияет на успех выполнения этапа поиска "описаны во внутренней библиотеке?".

Принципиальная схема работы механизма вывода оценок **LogP** приведена на рисунке 5.43:

5.10 Анализ большой СБД по ^{13}C спектрам

Предсказание спектра **^{13}C**, являющегося спектром резонансных частот атомов изотопа **^{13}C** [20], представляет значительный интерес из-за его широкого использования в химических исследованиях. Отличительной

особенность предсказания спектров химических структур от предсказания одного значения свойства является различие в структуре исходных данных. При построении **QSAR** -модели “структура-LogP” каждому соединению соответствовало одно число (значение LogP), соответственно модель должна была предсказывать только одно число для каждой новой структуры. Спектр **^{13}C** соединения представляет собой набор пар чисел: (сдвиг по спектральной шкале, интенсивность сигнала). Таким образом, каждой структуре в этом случае соответствует набор чисел - спектральных сдвигов.

Из теоретических соображений было принято решение прогнозировать не целиком спектральную картину для данной структуры, а выработать систему правил для прогнозирования спектральных сдвигов для отдельных атомов. Задача прогнозирования была облегчена решением отказаться от прогноза интенсивности сигнала, что в данном случае не приводит к существенной потери информативности спектральной картины. Ввод СБД и проведение расчетов выполнено совместно с аспирантами Митюшевым Д.Ф., Ощепковым Э.Н. и Колововановым Э.А.

В качестве исходных данных имелась база данных сдвигов **^{13}C** для более чем 25,000 структур. База данных представляла собой около 100,000 структурных фрагментов, представленных строкой символов, с приписанной величиной сдвига. Система кодирования фрагментов была модифицирована с учетом необходимости описывать стерические особенности структур. Был разработан программный инструментарий позволяющий проводить поиск подфрагментов из заданного списка на основе этой символьной кодировки. Экспертами были сформулированы классификационные признаки для разбиения базы данных на несколько частей по типу окружения атома с приведенным сдвигом. Для каждой из частей СБД был сформирован список фрагментов, теоретически вносящих аддитивный вклад в величину сдвига. Из таблицы 5.3 видно, что число фрагментов колеблется от 80 для алленов до 600 для ароматических углеводородов. Таблица с краткой характеристикой каждой из частей СБД (и соответствующей ей части БЗ) представлена в следующей таблице:

Таблица 5.3. Количество фрагментов в структурной библиотеке базы знаний "^{13}C", количество сдвигов в обучающей СБД и точность аппроксимации сдвигов из обучающей СБД по каждому из классов углеродных атомов.

расположения атома **C** в структуре.	Количество структурных фрагментов в библиотеке базы знаний	Количество эксперимен-тальных значений сдвигов	Точность аппроксимации: R^2	Стандартное отклонение предсказанного значения сдвига от эксперим. (**ppm**)
алифатический с 1-м соседом	351	37315	0.9798	2.2690
алифатический с 2мя соседами	351	10607	0.9922	2.3794
алифатический с 3мя соседами	298	3038	0.9999	2.6124
алифатический с 4мя соседями	279	744	0.9998	2.6275
алифатический кольцевой	182	1494	0.9695	2.5092
алкенный (с двойной связью)	416	7519	0.9722	2.5637
алкенный с **C=C** связью	81	94	0.9873	2.3033
алкенный с **C=O** связью	312	6876	0.9801	2.3284
алкенный с **C=N** связью	243	1188	0.9805	2.6624
алкенный с **C=S** связью	136	221	0.9801	2.3638
в цепи сопряженных двойных связей	244	1056	0.9970	1.6983
ароматический с двумя соседями	614	22861	0.9798	2.3134
ароматический с тремя соседями	118	777	0.9888	2.2065
алкинный (с тройной связью)	456	3126	0.9693	2.5681

В ходе расчетов выяснилось, что около 30% соединений (примерно 30000 сдвигов) описываются аддитивной схемой не достаточно точно (отклонения предсказанных значений сдвигов от экспериментальных составляют более 10 ppm). Полученные в результате нескольких итераций расчетов таблицы фрагментов и перекрестных членов (объем текстовых файлов с таблицами 275 Кб) составили основу базы знаний для прогнозирующего алгоритма.

После удаления из СБД, содержащей 120,000 сдвигов (100,000 из исходной базы и 20,000 введенных из других источников) около 10,000 фрагментов (т.е. 8% общего объема базы) предсказанный сдвиг стал

отличаться от экспериментального не более чем на 5 ppm, что является удовлетворительной точностью для данной задачи [348]. Принципиальная схема функционирования прогнозатора по базе знаний приведена на рисунке.

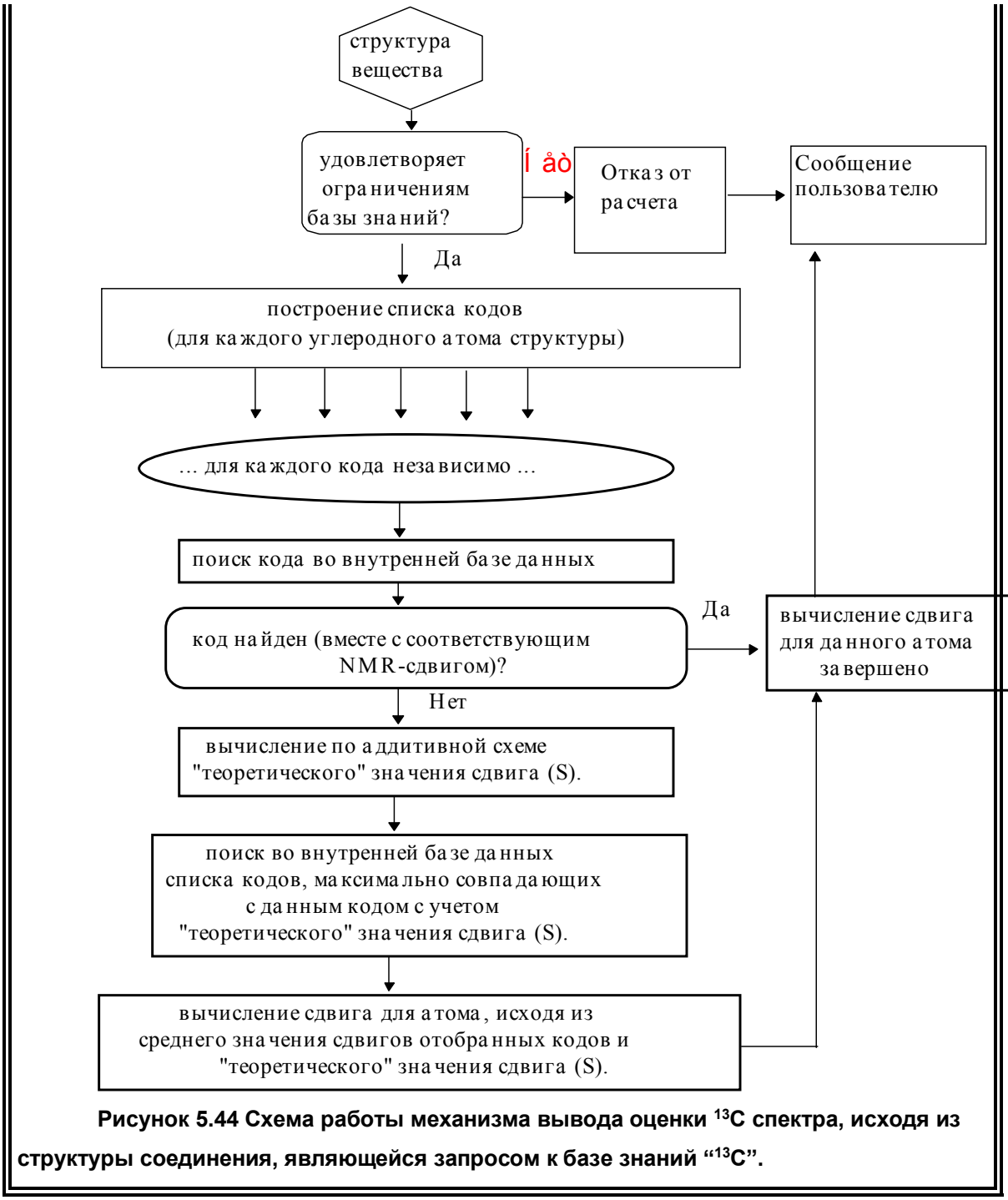


Рисунок 5.44 Схема работы механизма вывода оценки ^{13}C спектра, исходя из структуры соединения, являющейся запросом к базе знаний "^{13}C".

5.11 Выводы

Таким образом, проведена практическая проверка метода и разработанных программ для построения и селекции вариантов аддитивных схем расчета различных физико-химических свойств веществ на обучающих выборках органических соединений разных классов, включая гетероциклические соединения. Доказана прогностическая устойчивость найденных **QSAR**-зависимостей методом "скользящего контроля". Показано, что варьирование маркировки атомов и сложности структурных фрагментов существенно влияет на качество построенных **QSAR**-моделей. Показано, что приведенные лучшие модели действительно являются лучшими по критерию скользящий контроль ("cross-validation"). На построенных графиках зависимости $\mathbf{R^2}$ и $\mathbf{R^2_{cv}}$ видно, что увеличение числа дескрипторов в модели по сравнению с “лучшей моделью” не приводит к увеличению $\mathbf{R^2_{cv}}$, а в ряде случаев и ухудшает модель по этому критерию.

6. Глава 6. Прогнозирование биологической активности.

6.1 Использование трехмерного представления молекул

Äëÿ ïðåäñêàçàíèÿ áèîëîãè÷åñêèõ ñâîéñòâ âàæíóþ ðîëü èãðàåò ó÷åò ïðîñòðàíñòâåííîãî ñòðîåíèÿ ìîëåêóë. Äëÿ ïðîâåðêè ïðèìåíèìîñòè ïîèñûâàåìîé â ðàáîòå ìåòîäîëîãèè äëÿ àíàëèçà ïðîñòðàíñòâåííûõ îñîáåííîñòåé ìîëåêóëÿðíûõ ñòðóêòóð áûëè ïîñòðîåí ðÿä **QSAR**-ìîäåëåé äëÿ ïðîãíîçèðîâàíèÿ áèîëîãè÷åñêîé àêòèâíîñòè.

6.1.1 Антибактериальная активность триазинов.

Îáó÷àþùàÿ ÑÁÄ, ñîäåðæàùàÿ **54** ñîåäèíåíèÿ, áûëà ââåäåíà â ÝÂÌ íà Õèìôàêå ÌÃÓ, äàííûå î ñòðóêòóðàõ è áèîëîãè÷åñêîé àêòèâíîñòè áûëè âçÿòû èç ðàáîòû[349]. Ñâîéñòâîì ÿâëÿåòñÿ àíòèáàêòåðèàëüíàÿ àêòèâíîñòü - **MEC -** ìèíèìàëüíàÿ ýôôåêòèâíàÿ êîíöåíòðàöèÿ âåùåñòâà â ðàñòâîðå. Ñâîéñòâî ïåðåä ïîñòðîåíèåì **QSAR**-ìîäåëè áûëî ïðåîáðàçîâàíî ïî ôîðìóëå: $A_2=log(1/MEC)$.

Подготовка обучающих СБД к поиску **QSAR**-зависимостей состояла в расчете геометрии молекул с помощью программ молекулярной механики с использованием силового поля **MM2** с учетом внутримолекулярных водородных связей. Дистанции в формуле (**2.5**) строились, как правило, при следующем разбиении расстояний на интервалы:

Геометрические интервалы (в ангстремах **A°)** имели следующую нумерацию: **1**-(0-1.0); **2**-(1.0-1.2); **3**-(1.2-1.5); **4**-(1.5-2.0); **5**-(2.0-2.5); **6**-(2.5-3.0); **7**-(3.0-4.5); **8**-(4.5-5.0); **9**-(5.0-5.5); **10**-(6.0-∞).

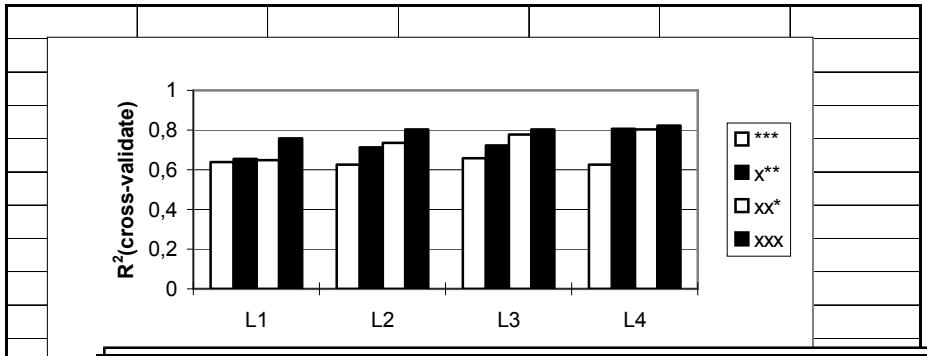
Нумерация *топологических интервалов*: **1**-(0-1); **2**-(1-2); **3**-(2-3); **4**-(3-4); **5**-(4-5); **6**-(5-6); **7**-(6-7); **8**-(7-8); **9**-(8-9); **10**-(9-∞).

Был проведен поиск **QSAR**-зависимостей на дескрипторах вида (2.5), когда примитивами являлись **1**-фрагменты (атомы) и **2**-фрагменты (пара “атом-атом”). Таким образом было сформировано **6** типов дескрипторов:

три "топологических" типа: **AAT**, **ABT**, **BBT**;

три "геометрических" типа: **AAG**, **ABG**, **BBG**;

Ñèìâîëû **G** è **T** óêàçûâàþò íà òèï èñïîëüçîâàííîãî ðàññòîÿíèÿ, (ãåîìåòðè÷åñêîãî èëè òîïîëîãè÷åñêîãî), à ñèìâîëû **A** è **B** - íà èñïîëüçîâàíèå â êà÷åñòâå áàçîâûõ ôðàãìåíòîâ àòîìîâ è ñâÿçåé ñîîòâåòñòâåííî. **QSAR** ðàñ÷åòû ïðîâîäèëèñü ïî êàæäîìó òèïó äåñêðèïòîðîâ îòäåëüíî. **QSAR**-ðàñ÷åòû ñ èñïîëüçîâàíèåì *òðåõìåðíûõ äåñêðèïòîðîâ* ïðîâåäåíû ñîâìåñòíî ñ Ìàêååâûì Ã.Ì. - àñïèðàíòîì Èíñòèòóòà Îðãàíè÷åñêîé Õèìèè ÐÀÍ.



Антикокцидная активность.

Рисунок 6.1 **Зависимость QSAR прогностичности от маркировки атомов и сложности примитива описания - длины цепочек атомов**

”Ñêîëüçÿùèé êîíòðîëü”: **N=54, K=10, R^2=0.821, F=20.18, S=0.481**.
Êîýôôèöèåíòû ëó÷øåé **QSAR**-ìîäåëè (êîäèðîâêà **XXX**, äëèíà öåïî÷åê **<= 4**):

1. -0.3978 (0.128) CONST
2. 0.4732 (0.136) F(X)= X C 1scS 2scC 3asC 3as
3. 0.3008 (0.013) F(X)= X C 2arC 2arC 3as
4. -0.8963 (0.039) F(X)= X C 2arC 2arC 3asN 3ss
5. 0.6959 (0.042) F(X)= X C 2tc
6. -1.4186 (0.394) F(X)= X C 3arC 2arC 3asS 2sc
7. 0.2627 (0.017) F(X)= X C 3asC 2arC 3asCL1sc
8. 0.2350 (0.013) F(X)= X F 1sc
9. -0.4702 (0.022) F(X)= X O 2sc
10. 0.5502 (0.040) F(X)= X S 2sc
11. 0.0578 (0.006) F(X)= X _LEN1

Наилучшие прогностические **QSAR**-модели были получены с использованием как *нелинейных преобразований переменных* при МГУА селекции уравнений, так и с включением в модель топологических индексов. При этом для получения устойчивой (прогностичной) модели потребовалось использование цепочек длины шесть с выключенным **p**-маркером.

Скользящий Контроль: **N=53, K=11, R^2_{cv}=0.895, S=0.38, F=31.88**

Коэффициенты лучшей **QSAR**-уравнения (маркировка: **“*XX”**,):

1. 2.801 свободный член
2. 0.214 цепочка F(X)=Sqrt(X) C *asC *asC *asC *as
3. -0.255 цепочка F(X)=Max-X C *asC *asC *asC *asC *as
4. -0.145 цепочка F(X)=Max-X C *asC *asC *asC *asC *asS *sc
5. 0.029 цепочка F(X)=X*X C *asC *asC *ssC *asC *asCL*sc
6. 0.353 цепочка F(X)=Sqrt(X) C *asC *asS *scC *asC *asC *ss
7. -0.055 цепочка F(X)=X*X C *asH *ss
8. 0.156 цепочка F(X)=X*X C *asN *asC *asC *asC *asCL*sc
9. -0.206 цепочка F(X)=Sqrt(X) C *ssC *asC *asS *sc
10. 0.436 цепочка F(X)= X H *ssC *asC *asC *asC *asH *ss
11. 0.127 цепочка F(X)= X H *ssC *asC *asC *asH *ss
12. 0.039 индекс F(X)=Max-X _T.diametr

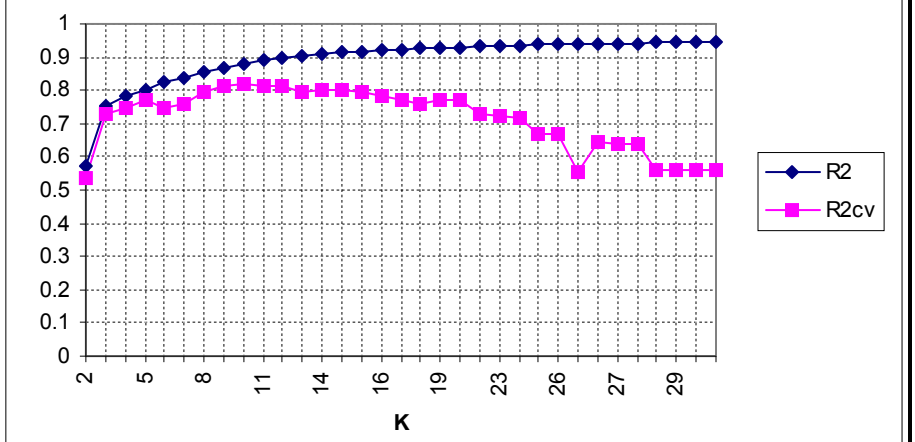


Рисунок 6.2

Антикокцидная активность. **Зависимость QSAR аппроксимации (R^2) и прогностичности (R^2cv) от сложности уравнения K - числа параметров (при заданной маркировке атомов)**

где **_T.diametr** - диаметр молекулярного графа, **Sqrt(X)** - преобразование "квадратный корень", **(Max-X)** - комплементарное число повторений цепочки **X**.

Скользящий Контроль: **N=53, K=11, R^2_{cv}=0.875, S=0.419, F = 25.73**

Коэффициенты **QSAR**-модели (кодировка "***XX**", длина цепочек **<= 4**):

1.	**0.832**	**свободный член**	
2.	**0.090**	**цепочка**	**F(X)= X**	**C *asC *asC *asC *as**
3.	**0.230**	**цепочка**	**F(X)= X**	**C *asC *asC *asS *dc**
4.	**-0.274**	**цепочка**	**F(X)=Max-X**	**C *asC *asC *asS *sc**
5.	**0.261**	**цепочка**	**F(X)= X**	**C *asC *asC *dcC *ss**
6.	**0.141**	**цепочка**	**F(X)= X**	**C *asC *asS *wcC *as**
7.	**-0.402**	**цепочка**	**F(X)=Max-X**	**C *dcC *asC *asCL*sc**
8.	**0.322**	**цепочка**	**F(X)= X*X**	**C *ssC *asC *asCL*sc**
9.	**0.080**	**цепочка**	**F(X)= X*X**	**CL*scC *asC *asH *ss**
10.	**0.230**	**цепочка**	**F(X)=Max-X**	**CL*scC *asC *asH *ss**
11.	**-0.037**	**цепочка**	**F(X)= X*X**	**H *ssC *asC *asH *ss**
12.	**0.238**	**цепочка**	**F(X)=Sqrt(X)**	**_R.len2**

где **_R.len.2** - топологический индекс, аналог индекса Рандича, вычисленный по цепочкам длины два, **X*X** - функциональное преобразование "возведение в квадрат".

Сравнение двух предыдущих моделей показывает, что на **6**-цепочках получена более устойчивая модель, чем на **4**-цепочках с той же маркировкой атомов (***XX**).

Скользящий Контроль: **N = 53, K = 11, R^2 = 0.861, S = 0.439, F = 23.14**

Коэффициенты **QSAR**-модели:

1. -8.301 свободный член
2. 0.015 цепочка F(X)= X*X C 3asC 3asC 3asC 3as
3. 0.328 цепочка F(X)= X C 3asC 3asC 3asC 4sc
4. -0.193 цепочка F(X)= X C 3asC 3asC 3asN 3dc
5. 0.662 цепочка F(X)= X*X C 4ssC 3asC 3asCL1sc
6. 0.928 цепочка F(X)=Max-X C 4ssC 3asC 3asCL1sc
7. -0.043 цепочка F(X)= X*X H 1ssC 3asC 3asH 1ss
8. -0.531 цепочка F(X)=Sqrt(X) H 1ssC 3asC 3asH 1ss
9. -0.990 цепочка F(X)=Sqrt(X) O 2sc
10. 0.088 цепочка F(X)= X S 2sc
11. -1.071 индекс F(X)=Sqrt(X) _R.len4
12. 4.951 индекс F(X)=Sqrt(X) _T.radius

Таким образом, на цепочках длины четыре модель с ***XX** маркировкой атомов более прогностична, чем на цепочках с полной **XXX-**маркировкой атомов.

Таблица 6.1 **Результаты построения** QSAR**-зависимостей на трехмерных дескрипторах (число параметров** K=10**) для антикокцидной активности. Коррелируемый параметр - логарифм минимальной эффективной концентрации. Приведены модели с лучшим значением** R^2_{cv}.

Тип дескрипторов	R^2	R^2_{cv}
AAG	0.890	**0.857**
ABG	0.895	0.856
BBG	0.892	0.844
AAT	0.899	**0.843**
ABT	0.872	0.784
BBT	0.897	0.831

Полученные результаты сопоставимы с известными литературными данными. В работе [349] McFarland получил величину **R^2=0.74 (R=0.86)** при использовании CoMFA метода. В работе [350] на 7 дескрипторах получено значение **R^2=0.81 (R=0.9)** для модели аппроксимации. Использование нами систематического поиска дескрипторов, адекватных для исследуемой активности, в совокупности с применением эволюционного алгоритма позволило получить более высокие значения R^2 не на аппроксимации, а *на тестировании методом "скользящего контроля*". Полученные результаты свидетельствуют, что данное биологическое свойство хорошо описывается *уже на топологическом уровне*: переход к геометрическим дескрипторам не привел к существенному улучшению качества модели.

6.1.2 Анти-ВИЧ активность.

Обучающая СБД была введена в ЭВМ Г.М.Макеевым по работе [351] и содержала **45** полигидроксипиперидинов и полигидроксипирролидинов, **19** из которых проявляют анти-ВИЧ активность. Проявляемая ими способность подавлять развитие вируса ВИЧ обусловлена ингибированием α-глюкозидазы вируса и существенно зависит от пространственного строения соединения. В качестве активности рассматривалась процентная доля ингибирования репликации вируса в культурах клеток in vitro. Анти-ВИЧ активность (% подавления репликации вируса) аппроксимировалась к единой концентрации по методике, описанной в [352]. В молекулах обучающей СБД (после расчета программами молекулярной механики) определялись *абсолютные конфигурации*. Тем атомам углерода, которые обладали абсолютной конфигурацией **R** и **S**, приписывался ***маркер*** **CR** и **CS** соответственно.

Топологические дескрипторы вида **AAT** и **ABT,** как и ожидалось, обладают слабой прогностичной способностью. Неожиданным результатом являются высокие прогностические качества модели, содержащей всего четырех **топологических** дескриптора вида **BBT** (Таблица 6.4)**.** Таблица 6.2 Результаты для анти-ВИЧ активности на геометрических дескрипторах. Таблица 6.3 Вид дескрипторов QSAR-уравнения, построенного на геометрических дескрипторах BBG.

Таблица 6.2 Результаты для анти-ВИЧ активности на ***геометрических*** дескрипторах.

K-число дескрипторов	**Дескрипторы типа AAG**		**Дескрипторы типа ABG**		**Дескрипторы типа BBG**	
	R^2	R^2cv	R^2	R^2cv	R^2	R^2cv
1	**0.41**	**0.34**	**0.62**	**0.58**	**0.51**	**0.40**
2	**0.56**	**0.35**	**0.71**	**0.63**	**0.77**	**0.72**
3	**0.63**	**0.45**	**0.76**	**0.69**	**0.82**	**0.79**
4	**0.70**	**0.48**	**0.79**	**0.71**	**0.82**	**0.79**
5	**0.74**	**0.57**	**0.81**	**0.75**	**0.86**	**0.77**
6	***0.77***	***0.61****	**0.82**	**0.75**	**0.87**	**0.82**
7	**0.80**	**0.64**	**0.84**	**0.76**	***0.88***	***0.82****
8	**0.82**	**0.63**	***0.86***	***0.80****	**0.89**	**0.76**

9	**0.84**	**0.66**	**0.87**	**0.80**	**0.90**	**0.80**

Таблица 6.3 Вид дескрипторов QSAR-уравнения, построенного на геометрических дескрипторах BBG.

Номер i	**b_i**	**f_1 фрагмент**	**f_2 фрагмент**	**G интервал**
0	**11.978**	**CONST**		
1	**93.690**	**$-CR_{(r)}-C(c)-$**	**$-C_{(c)}-C(c)-$**	**$2.0 \le D_g < 2.5$**
2	**-12.973**	**$-C_{(c)}-H$**	**$-CR_{(r)}-C(r)-$**	**$2.5 \le D_g < 3.0$**
3	**8.955**	**$-CR_{(r)}-H$**	**$-C_{(c)}-H$**	**$1.5 \le D_g < 2.0$**
4	**-4.083**	**$-CR_{(r)}-H$**	**$-C_{(r)}-H$**	**$2.5 \le D_g < 3.0$**
5	**88.994**	**$-C_{(r)}-N<$**	**$-CR_{(r)}-CR(r)-$**	**$2.5 \le D_g < 3.0$**
6	**17.240**	**$-C_{(c)}-O-$**	**$-CR_{(r)}-CR(r)-$**	**$2.5 \le D_g < 3.0$**
7	**-3.083**	**H-O-**	**$-CR_{(r)}-H$**	**$2.0 \le D_g < 2.5$**

Обозначения: **b_i** - весовой коэффициент **i**-ого дескриптора, описывающего пару фрагментов **(f1,f2)**, расположенных друг от другая в молекуле в заданном интервале расстояний **G**; **CR** - атом углерода с **R** абсолютной конфигурацией; подстрочные индексы: **(r)** - атом находится в кольце, **(c)** - атом находится в цепи.

Таблица 6.4 **Результаты для анти-ВИЧ активности на** топологических **дескрипторах.**

K число дескрипторов	**Дескрипторы типа AAT**		**Дескрипторы типа ABT**		**Дескрипторы типа BBT**	
	R^2	***R^2_{CV}***	***R^2***	***R^2_{CV}***	***R^2***	***R^2_{CV}***
1	**0.54**	**0.44**	**0.56**	**0.51**	**0.56**	**0.51**
2	**0.66**	**0.45**	**0.64**	**0.50**	**0.78**	**0.76**
3	**0.70**	**0.47**	**0.68**	**0.59**	**0.83**	**0.81**
4	**0.73**	**0.52**	**0.71**	**0.51**	***0.85***	***0.83****
5	**0.76**	**0.52**	**0.74**	**0.57**	**0.87**	**0.82**
6	**0.78**	**0.55***	**0.77**	**0.62***	**0.89**	**0.83**
7	**0.79**	**0.51**	**0.78**	**0.60**	**0.89**	**0.83**
8	**0.81**	**0.49**	**0.80**	**0.52**	**0.90**	**0.83**
9	**0.82**	**0.48**	**0.82**	**0.36**	**0.91**	**0.84**

Было проведено сравнение геометрических и топологических дескрипторов, описывающих пары фрагментов, с базовыми дескрипторами системы **BIBIGON**, которые описывают топологические характеристики молекул в виде цепочек маркированных атомов. На **1-9** дескрипторах системы **BIBIGON** значение квадрата коэффициента корреляции ($\mathbf{R^2}$) и результаты скользящего контроля ($\mathbf{R^2_{CV}}$)не превышали значений $\mathbf{R^2 \leq 0.88}$ и $\mathbf{R^2_{CV} \leq 0.71}$. Таким образом, рассмотренные в разделе структурные дескрипторы, описывающие несвязные подструктуры, состоящие из двух фрагментов, имеют существенное преимущество для получения прогностически устойчивых **QSAR**-моделей.

Полученные результаты вычислительных экспериментов позволяют сделать важный вывод. При поиске адекватного описания молекул следует *последовательно наращивать уровень сложности* структурного спектра. Так, "топологические" дескрипторы **BBT** в наших расчетах, которые с точки зрения вычислительных затрат на порядок проще геометрических дескрипторов, позволили получить прогностически лучшие **QSAR**-уравнения с меньшим числом параметров, по сравнению с "геометрическими" дескрипторами.

6.2 Использование поверхности электростатического потенциала для формирования структурных дескрипторов

Биологическая активность определяется способностью молекулы комплементарно фиксироваться и взаимодействовать с функциональной химической структурой (например, белком) - “рецептором” [353]. Считается, что для комплементарности молекулы и рецептора необходимо соответствие как геометрических форм молекулы и рецептора [354,355,356], так и комплементарность электростатических полей молекулы и структуры рецептора [357]. В работах [358,359] отмечено, что важным фактором, влияющим на биологическую активность молекул, является электростатическое поле, создаваемое молекулой, и комплементарность этого поля структуре рецептора. Комплементарность формы молекулы и рецептора означает, что система “молекула-рецептор” должна подчиняться определенным геометрическим ограничениям: координирующаяся молекула должна находиться в тесном контакте с молекулой рецептора, расстояние между центрами атомов мо-

лекулы и рецептора не могут быть меньше суммы Ван-дер-Ваальсовых радиусов контактирующих атомов, что означает отсутствие химической модификации молекулы [360].

Поверхность электростатического потенциала. Очевидно, что любая молекула может быть окружена такой поверхностью, что центры атомов любой другой молекулы не могут находиться внутри этой поверхности, если между молекулами не происходит образования новых химических связей. То есть, именно такая поверхность отражает геометрические ограничения, накладываемые на систему “молекула-рецептор”: имеется тесный контакт Ван-дер-Ваальсовых сфер и отсутствует химическая модификация молекулы биологически активного вещества. Назовем такую поверхность **“покрывающей поверхностью”.**

В то время как геометрические факторы действуют на малых расстояниях порядка суммы Ван-дер-Ваальсовых радиусов и проявляют себя как жесткие геометрические ограничения, электростатически молекулы начинают взаимодействовать между собой на больших расстояниях, и характер этого взаимодействия изменяется плавным образом по мере сближения взаимодействующих молекул.

Энергия электростатического взаимодействия молекулы и рецептора увеличивается по мере их сближения, поэтому на минимальном расстоянии

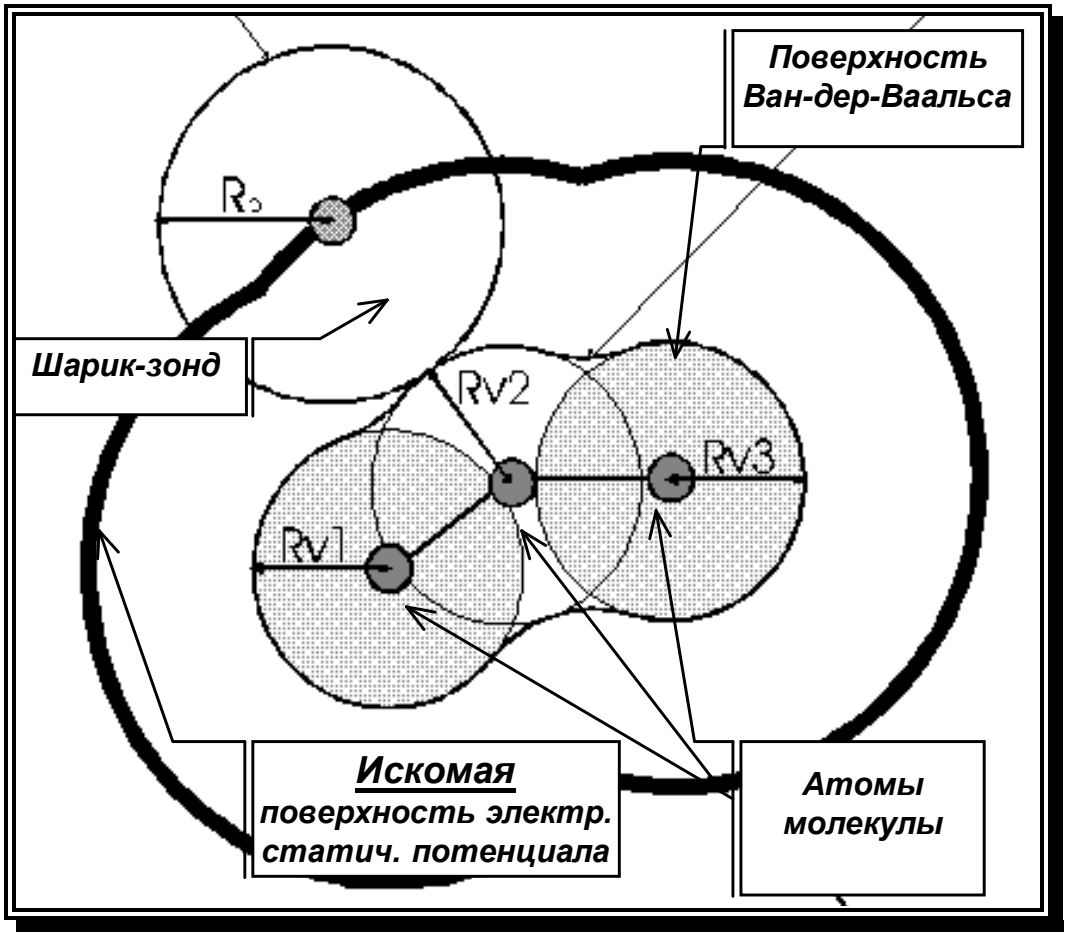


Рисунок 6.3 **Схема вычисления поверхности электростатического потенциала для модельной трехатомной молекулы.**

между ними взаимодействие будет максимальным. Поскольку “покрывающая поверхность” является геометрическим местом точек, максимально близко расположенных к молекуле и таких, что в них могут располагаться центры атомов другой молекулы (рецептора), то описание характеристик электростатического поля на этой поверхности позволит учесть геометрические и электростатические факторы.

Поставим в соответствие каждой точке “покрывающей поверхности” величину молекулярного электростатического потенциала и назовем полученную таким образом поверхность **“поверхностью электростатического потенциала”** [361]. Анализируя форму этой поверхности и характер распределения молекулярного электростатического потенциала на ней, можно выделить закономерности, общие для набора молекул, обладающих одним и тем же видом биологической активности на основе комплементарности форм молекулы и рецептора [362]. “Поверхность электростатического потенциала” была определена как покрывающая поверхность с заданными на ней значениями МЭП. (Покрывающая поверхность -- это геометрическое место точек, созданных центром шарика-зонда при облегающем прокатывании его по поверхности Ван-дер-Ваальса). Путем проведения ряда вычислительных экспериментов было найдено, что оптимальное значение радиуса шарика зонда лежит в пределах **1.4-1.6A**. Для формирования покрывающих поверхностей была использована величина радиуса **1.5A**.

6.2.1 Общая схема формирования дескрипторов, описывающих особенности пространственной формы молекул .

1. Молекула рассматривается как пространственный объект, обладающий определенной структурой, задаваемой пространственными координатами атомов и таблицей связи атомов.

2. На основе экспертных оценок производится выделение в молекуле или ее окрестностях "особых точек" или "точек представления".

3. Для каждой особой точки определяются пространственные координаты и задается тип. Тип точки определяется на основании правил, устанавливаемых экспертом.

4. Для всех особых точек молекулы строится матрица расстояний **D**=$\{d_{IJ}\}$, где d_{IJ} - евклидово расстояние между i-ой и j-ой точками. На основе экспертной оценки (или путем вычислительного эксперимента) выбирается разбиение расстояний на интервалы и строится “интервальная” матрица **P**=$\{p_{IJ}\}$, где p_{IJ} - номер интервала, в который попало расстояние d_{IJ}.

5. Полученная интервальная матрица является основой для построения описания объекта путем полного перечисления всех пар особых точек. Формируется список повторений всех пар точек в виде $(T_1,T_2,P),N$, где: T_1 и T_2 -

типы (имена) особых точек, входящих в пару, P - номер интервала расстояния между ними, N - число повторений фрагмента (T_1,T_2,P) в объекте. При этом считается, что символьная запись (T_1,T_2,P) есть имя дескриптора. *Дескрипторы равны, если совпадают их имена*. Особые точки и номера интервалов расстояний перечисляются в коде дескриптора в возрастающем лексикографическом порядке.

6. Пары особых точек перечисляются в каждой молекуле обучающей СБД. Составляется общий список различных признаков, найденных в данной СБД, и на его основе формируется матрица описания "молекула-признак" **X**={ x_{IJ} }, где x_{IJ} - число повторений j-ого признака в i-ой молекуле.

Далее в системе BIBIGON проводится отбор значимых признаков и строится семейство моделей, наиболее адекватно описывающих заданное свойство в терминах типов особых точек, предложенных экспертом.

7. На основе полученных **QSAR** -моделей и соответствующего набора признаков, входящих в модели, выявляется физическое содержание особенностей формы поверхности и пространственной структуры молекул, ответственных за проявление заданного свойства.

Пример применения методики. Продемонстрируем на условном примере этапы 1-5 общей схемы, направленные на формирование списка дескрипторов и получение описания молекулы в виде структурного спектра.

Шаг 1: На рисунке показана условная "плоская" молекула, состоящая из трех атомов "A", "B", "C". Координаты центров этих атомов будут: X_A=1.0,Y_A=1.0; X_B=2.0,Y_B=2.0; X_C=3.4,Y_C=2.0 (координаты приведены в условных единицах).

Вокруг каждого атома опишем сферу с радиусом, равным радиусу Ван-дер-Ваальса. Пусть для всех трех типов атомов эти радиусы будут равны 1.5 (условной величины): R_A=R_B=R_C=1.5. Допустим, что в результате компьютерных расчетов были получены величины зарядов на атомах: Q_A=+0.25, Q_B=-0.5, Q_C=+0.25.

Шаг 2: Выберем точки представления двумя способами - а)точки на ПЭП; б) точки на центрах атомов.

Шаг 2а (точки на ПЭП): Построим вокруг молекулы покрывающую поверхность (показана тонкой линией на расстоянии R_p от поверхности Ван-дер-

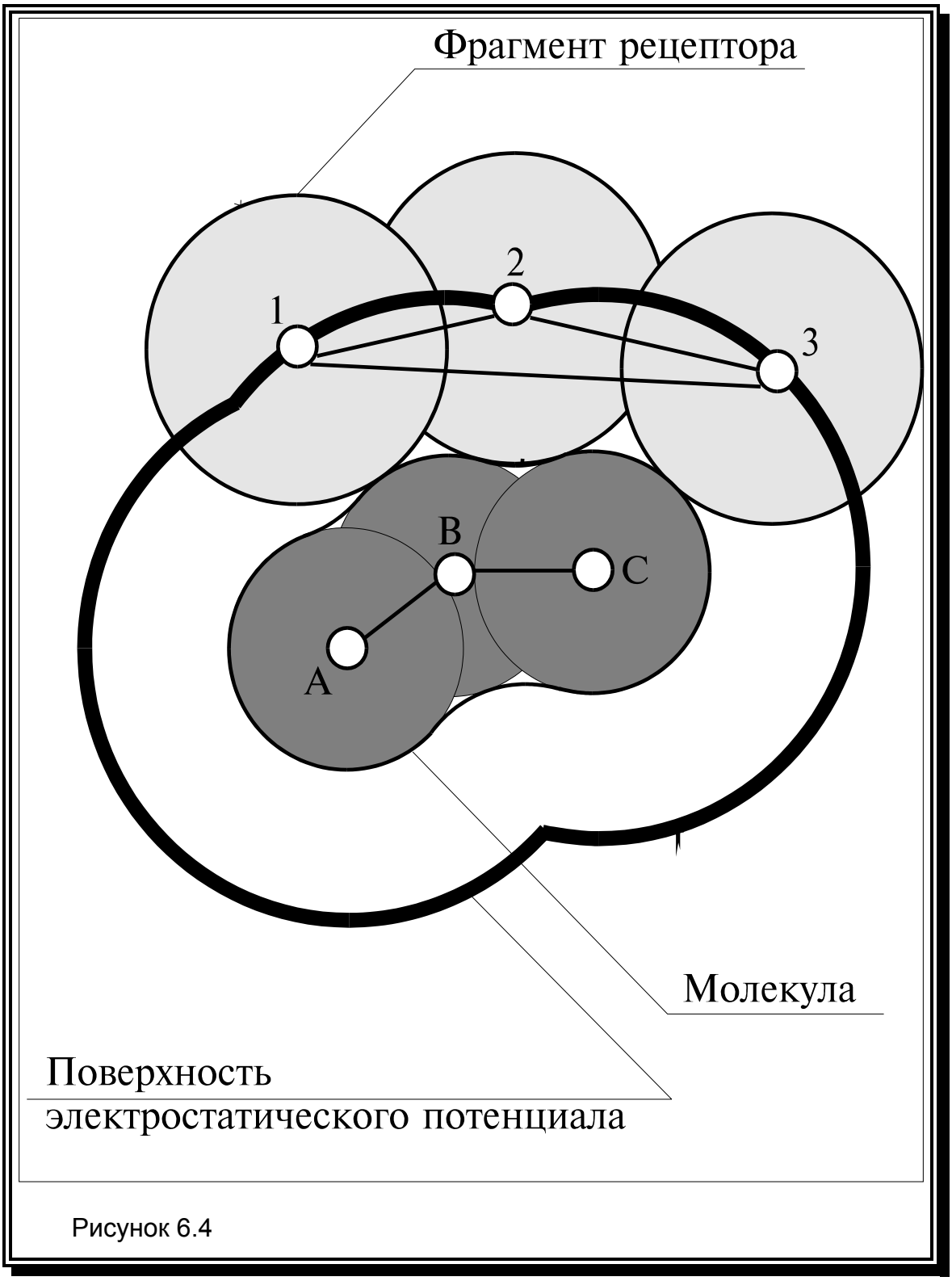


Рисунок 6.4

Ваальса). Выберем значение R_p=1.5 (что будет в нашем примере соответствовать среднему значению радиуса Ван-дер-Ваальса для атомов, входящих в состав молекул органических веществ). Покроем полученную поверхность точками со средним шагом между ними 0.3 (в условных единицах). Затем вычислим значения молекулярного электростатического потенциала (МЭП) в этих точках. Найдем теперь такие точки, значение МЭП в которых достигает локального максимума или минимума (например, в случае локального максимума в точке 3, ее ближайшие соседи - точки 2 и 4 - будут иметь значение потенциала меньшее, чем значение потенциала в точке 3), - эти точки и будем в дальнейшем использовать в качестве точек представления.

Шаг 3а (точки на ПЭП): Предположим, что в нашем случае в качестве точек максимума были выявлены точки α, β, γ, а в качестве точки минимума выявлена точка δ. Допустим что координаты этих точек имеют следующие значения: X_α=-2.0,Y_α=-2.0; X_β=3.2,Y_β=-0.7; X_γ=6.4,Y_γ=2.0; X_δ=0.8,Y_δ=4.8.

Пусть значения МЭП в этих точках (выраженные в условных единицах) будут: U_α=+6.0, U_β=+2.0, U_γ=+6.0, U_δ=-2.0.

Разобьем весь диапазон значений МЭП на ряд интервалов: $E_1\equiv(-\infty,0.0)$, $E_2\equiv(0.0,+3.0)$, $E_3\equiv(+3.0,+\infty)$. Тогда значения МЭП в точках α и γ попадут в интервал E_3, значения МЭП в точке β попадет в интервал E_2, а значение МЭП в точке δ попадет в интервал E_1. Теперь зададим типы точек, имена которых совпадают с именами интервалов, и припишем каждой точке ее тип: $\alpha\rightarrow E_3$, $\beta\rightarrow E_2$, $\gamma\rightarrow E_3$, $\delta\rightarrow E_1$.

Шаг 4а (точки на ПЭП): Вычислим расстояния между точками: $L_{\alpha\beta}$=5.36 (в условных единицах), $L_{\alpha\gamma}$=9.30, $L_{\alpha\delta}$=7.28, $L_{\beta\gamma}$=4.19, $L_{\beta\delta}$=6.02, $L_{\gamma\delta}$=6.19.

Разобьем весь диапазон значений расстояний на несколько интервалов: $\lambda_1\equiv(3.0,4.5)$, $\lambda_2\equiv(4.5,6.5)$, $\lambda_3\equiv(6.5,8.5)$, $\lambda_4\equiv(8.5,10.0)$. Тогда $L_{\alpha\beta}$, $L_{\beta\delta}$, $L_{\gamma\delta}$ попадут в интервал λ_2; $L_{\alpha\gamma}$ попадет в интервал λ_4; $L_{\alpha\delta}$ попадет в интервал λ_3; а $L_{\beta\gamma}$ попадет в интервал λ_1.

Шаг 5а (точки на ПЭП): Формируем список дескрипторов вида (T_1,T_2,P):

D_1: (E_3,E_2,λ_2), для пары точек α,β;

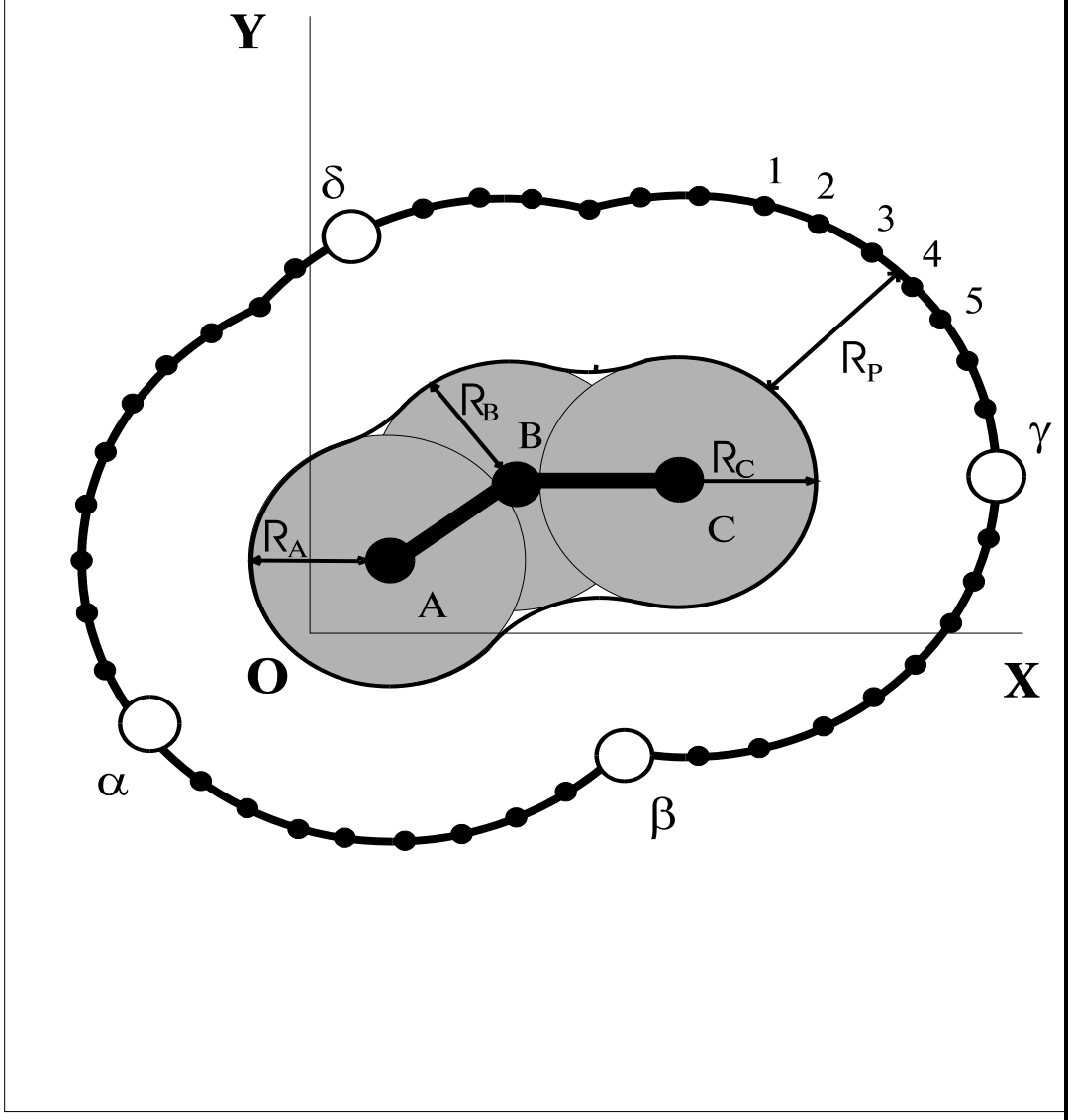


Рисунок 6.5

D_2: (E_3,E_3,λ_4), для пары точек α,γ;
D_3: (E_3,E_1,λ_2), для пары точек α,δ;
D_4: (E_2,E_3,λ_1), для пары точек β,γ;
D_5: (E_2,E_1,λ_2), для пары точек β,δ;
D_6: (E_3,E_1,λ_2), для пары точек γ,δ.

После упорядочения типов точек в дескрипторе в соответствии с возрастающим лексикографическим порядком будем иметь:

D_1: (E_2,E_3,λ_2), для пары точек α,β;
D_2: (E_3,E_3,λ_4), для пары точек α,γ;
D_3: (E_1,E_3,λ_2), для пары точек α,δ;

D_4: (E_2,E_3,λ_1), для пары точек β,γ;
D_5: (E_1,E_2,λ_2), для пары точек β,δ;
D_6: (E_1,E_3,λ_2), для пары точек γ,δ.

Заметим, что D_3 и D_6 полностью идентичны. Сформируем завершающий список дескрипторов вида (T_1,T_2,P),N:

D_1': (E_2,E_3,λ_2),1
D_2': (E_3,E_3,λ_4),1
D_3': (E_1,E_3,λ_2),2
D_4': (E_2,E_3,λ_1),1
D_5': (E_1,E_2,λ_2),1

Этот список представляет собой структурный спектр, построенный на точках, расположенных на ПЭП.

Шаг 2б (точки на центрах атомов): В качестве точек представления используем центры атомов, т.е., имеем три точки: A,B,C.

Шаг 3б (точки на центрах атомов): Как указывалось выше, точкам соответствуют следующие значения координат: X_A=1.0,Y_A=1.0; X_B=2.0,Y_B=2.0; X_C=3.4,Y_C=2.0 (координаты приведены в условных единицах).

Будем использовать для задания типа точек рассчитанные ранее значения заряда на атомах. Тогда мы имеем точки двух типов: Q_1 - точки A и C, Q_2 - точка B. При этом мы полагаем, что тип Q_1 приписывается точкам, значение заряда на которых положительно; тип Q_2 приписывается точкам, значения заряда на которых отрицательно.

Шаг 4б (точки на центрах атомов): Вычислим расстояния между точками: L_{AB}=1.4, L_{BC}=1.4, L_{AC}=2.6 (в условных единицах). Разобьем весь диапазон расстояний на два интервала: $\lambda_1\equiv(0.0,2.0)$, $\lambda_2\equiv(2.0,+\infty)$. Тогда L_{AB}, L_{BC} попадут в интервал λ_1, а L_{AC} попадет в интервал λ_2.

Шаг 5б (точки на центрах атомов): Формируем дескрипторы вида (T_1,T_2,P):

D_1: (Q_1,Q_2,λ_1), для точек A,B;
D_2: (Q_2,Q_1,λ_1), для точек B,C;
D_3: (Q_1,Q_1,λ_2), для точек A,C.

После проведения упорядочения типов точек в дескрипторе в соответствии с возрастающим лексикографическим порядком будем иметь:

D_1: (Q_1,Q_2,λ_1), для точек A,B;
D_2: (Q_1,Q_2,λ_1), для точек B,C;
D_3: (Q_1,Q_1,λ_2), для точек A,C.

Заметим, что D_1 и D_2 полностью идентичны. Сформируем завершающий список дескрипторов вида (T_1,T_2,P),N:

D_1': (Q_1,Q_2,λ_1),2
D_2': (Q_1,Q_2,λ_1),1

Этот список представляет собой структурный спектр, построенный на точках, расположенных на центрах атомов.

Ниже приведены результаты анализа двух видов биологической активности: ***запаха ароматических мускусных одорантов*** [363] и ***психотропной активности азотсодержащих гетероциклических соединений*** [364].

Оптимизация геометрии выполнялась в два этапа: *начальная* оптимизация - методами молекулярной механики с использованием силового поля *MMX*[365]; *конечная* оптимизация - методами квантовой химии с использованием гамильтониана *AM1*[366] при помощи программы MOPAC [367]. Для построения дескрипторов были исследованы **6** типов особых точек, с каждым из которых производилась своя серия вычислений:

•атомы молекулы с приписанными им именами (*серия вычислений **I***);

•центры атомов молекулы с приписанными им величинами зарядов без имен элементов. Использовались заряды атомных остовов и формальные заряды (т.е. такие условные заряды на атомах, на основании которых электростатическое поле, создаваемое молекулой, может быть аппроксимировано наилучшим образом) (*серии вычислений **II,IV***);

•центры атомов с приписанными им именами и величинами зарядов. (заряды атомных остовов и формальные заряды (*серии вычислений **III,V**)*);

•точки на “поверхности электростатического потенциала”, в которых молекулярный электростатический потенциал (МЭП) имеет локальные экстремумы; каждой точке приписано значение МЭП (*серия вычислений **VI***).

Ëèàïàçîíû çíà÷åíèé çàðÿäîâ àòîìíûõ îñòîâîâ è ôîðìàëüíûõ çàðÿäîâ áûëè ðàçäåëåíû íà èíòåðâàëû â çàâèñèìîñòè îò ÷àñòîòû âñòðå÷àåìîñòè ýòèõ âåëè÷èí â îáó÷àþùåé âûáîðêå. *Ëèàïàçîíû çíà÷åíèé ðàññòîÿíèé* ìåæäó îñíîâûìè òî÷êàìè áûëè ðàçäåëåíû íà èíòåðâàëû ñî ñðåäíèì øàãîì **0.2À**.

Ïðè êëàññèôèêàöèè ïîëàãàëîñü, ÷òî âåùåñòâî íàëàäàåò çàïàõîì, åñëè **Ñ>0.5**, â ïðîòèâíîì ñëó÷àå ïîëàãàëîñü îòñóòñòâèå çàïàõà. **QSAR**-ðàñ÷åòû ñ èñïîëüçîâàíèåì ïîâåðõíîñòü ýëåêòðîñòàòè÷åñêîãî ïîòåíöèàëà ïðîâåäåíû ñîâìåñòíî ñ Çûðÿíîâûì È.Ë. - àñïèðàíòîì Èíñòèòóòà Îðãàíè÷åñêîé Õèìèè ÐÀÍ è Ñâèòàíüêî È.Â. - äåêàíîì Âûñøåãî õèì÷åñêîãî êîëëåäæà.

6.2.2 Мускусный запах.

Обучающая СБД была введена в ЭВМ аспирантом ИОХ РАН И.Л.Зыряновым и содержала **110** соединений (рис. 5.3), из которых 35 обладают ярко выраженным мускусным запахом, а остальные лишены его.

Мускусный запах выбран потому, что он не является тривиальным аддитивным свойством (“аддитивное” свойство - это свойство, величина которого, может быть с определенной погрешностью представлена в виде суммы “инкрементов” отдельных структурных элементов; как следствие этого, свойство изменяется на вполне конкретную величину при определенном изменении в структуре соединения) и в недостаточной степени прогнозирутся простыми аддитивными схемами. Кроме того, при анализе запаха мы можем не принимать во внимание транспорт и превращение молекул биологически активных веществ в организме, что позволяет сконцентрировать внимание непосредственно на комплементарности молекулы и рецептора.

Исследование мускусного запаха было предпринято ранее [368], и были предложены модели, описывающие мускусный запах органических соединений. Согласно [368], для того, чтобы вещество обладало мускусным запахом необходимо наличие в его структуре двух пар групп:

1) кислородсодержащая группировка **“>X=0” (-NO_2,-C(O)R)** и четвертичный атом углерода, находящийся на расстоянии **5.9-6.2A** от атома кислорода;
2) две метильные (или метиленовые) группы на расстоянии **5.3-5.8A**.

Данные признаки легко позволяют установить отсутствие запаха у соединения -- отсутствие какой-либо пары групп в молекуле приводит к полному исчезновению запаха. В то же время, данная модель не предлагает достаточных условий наличия мускусного запаха.

Лучшая из моделей, приведенных на рис 5.2 (кривая “VI”) была сопоставлена с лучшими моделями, построенными на дескрипторах системы БИБИГОН. Результаты сопоставления приведены на рис. 5.5. Приведенные данные показывают, что **QSAR**-модели (кривые “VI”, “VI+”), построенные примитивах, не связанных с атомами, но описывающих особенности формы

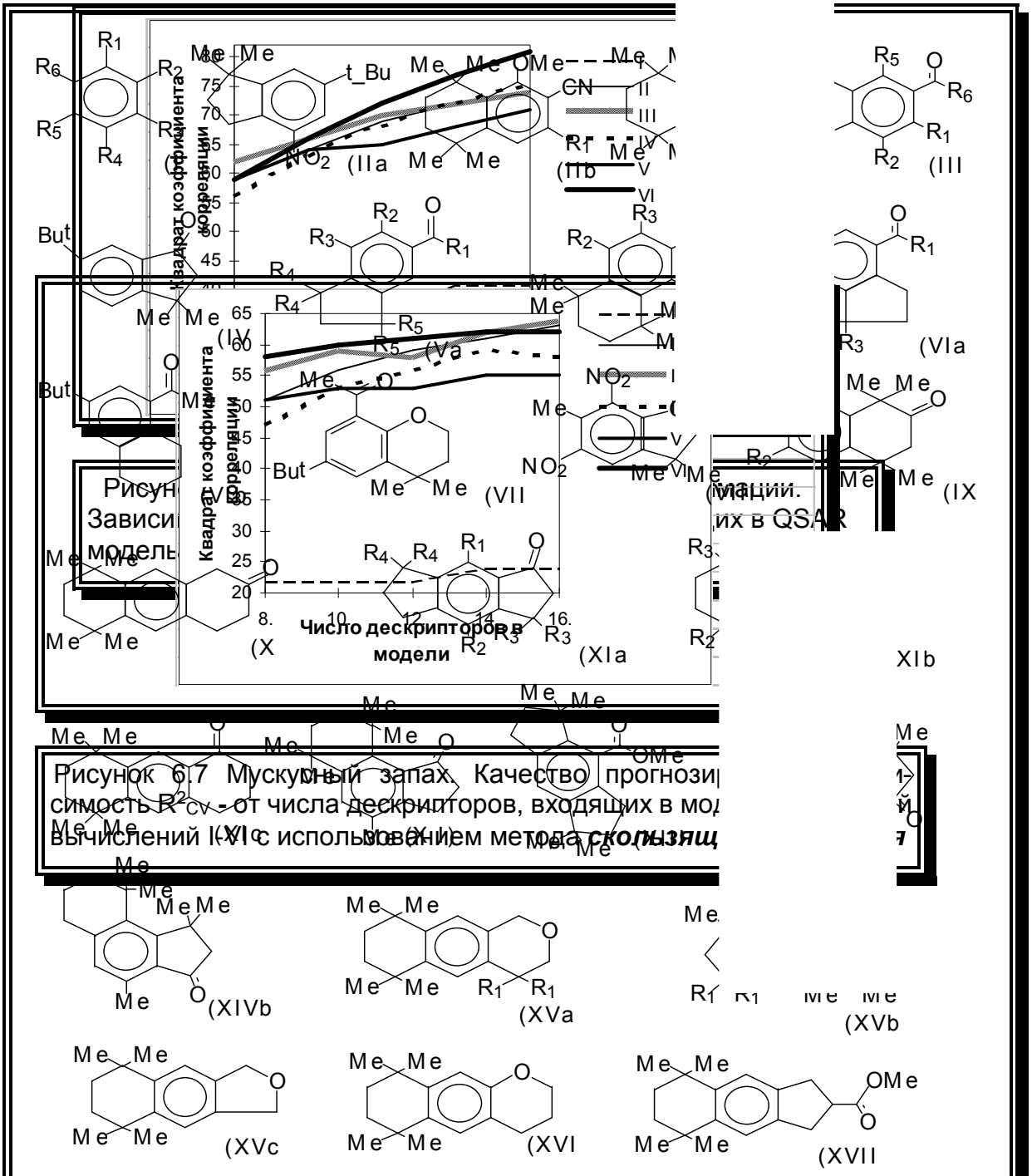


Рисунок 6.7 Мускусный запах. Качество прогнози… симость R^2_{CV} - от числа дескрипторов, входящих в мо… вычислений I-VI с использованием метода ***скользящ…***

Рисунок 6.8 **Базовые структуры мускусных одорантов.**

поверхности молекулы и распределение молекулярного электростатического потенциала на ней, обладают большей прогностической способностью, чем **QSAR**-модели системы БИБИГОН.

Использование трехмерного представления молекулы с расчетом поверхности электростатического потенциала позволяет сформировать достаточно адекватное описание для исследуемого вида активности.

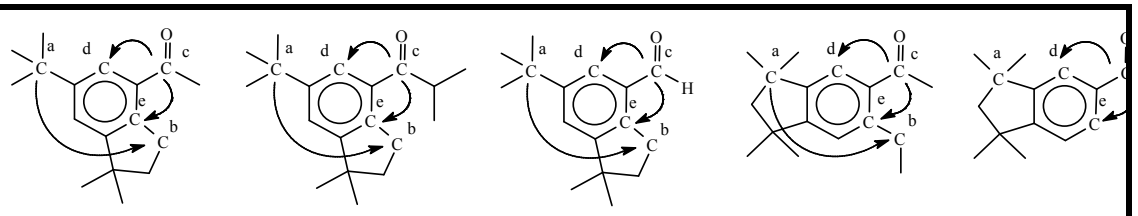


Рисунок 6.9 **Мускусный запах. Расположение систем особых точек, отвечающих за наличие мускусного запаха.** Act=1 **- вещество обладает запахом,** Act=0 **- вещество не обладает запахом;** a,b,c,d,e **- особые точки, расположенные на центрах атомов и входящие в состав систем особых точек; в качестве метки использован формальный заряд.** (ab),(cd),(ce) **- расстояния между особыми точками. В круглых скобках указаны допустимые пределы формального заряда и межточечных расстояний. ** - указывает, что данный параметр сильно выходит за допустимые пределы.**

На основании полученных **QSAR**-моделей были найдены такие сочетания особых точек, которые присутствуют в структурах веществ, обладающих мускусным запахом и отсутствуют в структурах, лишенных его. Также была найдена система особых точек, расположенных на поверхности электростатического потенциала, которая не встречается в молекулах, не обладающих запахом, но присутствует в тех структурах, которые им обладают. Эта система представляет собой треугольник с следующими параметрами: a=(**-510, -180**)Дж/моль, b=(**-690, -360**)Дж/моль, c=(**+220, +350**)Дж/моль; (ab)=(**3.85, 5.90**)A, (ab)=(**6.2, 7.4**)A, (ac)=(**7.0, 8.0**)A.

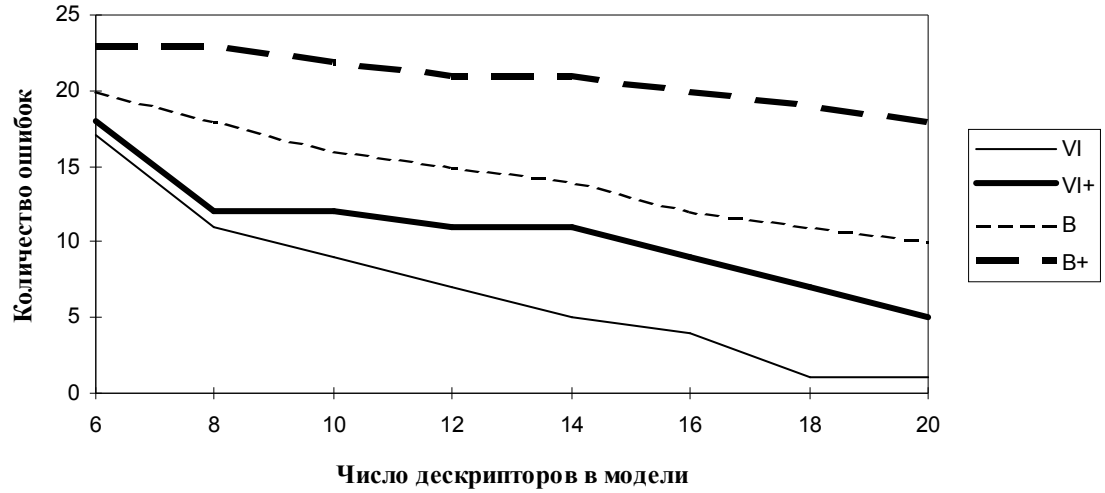


Рисунок 6.10 **Мускусный запах. Зависимость между количеством ошибок классификации и числом дескрипторов в модели.** "VI" **- кривая, построенная для серии VI,** "B" **- кривая, построенная для** QSAR**-моделей системы БИБИГОН.** "VI+" **и** "B+" **-кривые для тех же моделей, но построенные с использованием метода *скользящего контроля*.**

6.2.3 Психотропная активность.

Психотропная активность азотсодержащих гетероциклов была выбрана для исследования, поскольку эти вещества представляют в большинстве своем конформационно **относительно жесткие** структуры, что в значительной степени снижает разнообразие форм их молекул и повышает точность их описания. Кроме того, соединения, входящие в обучающую СБД представляют собой неметаболизирующие (т.е. не вступающие в химические превращения) вещества. Это позволяет предположить, что наибольшее влияние на активность этих соединений оказывает именно координация их молекул на рецепторе. Была использована обучающая СБД, содержащая **47** соединений, из которых **9** обладают ярко выраженной активностью, **26** соединений неактивны, а оставшиеся **12** соединений обладают промежуточной активностью (рис 5.8). Соединения были разделены на три класса: класс "1" - высокая активность, класс "0" - низкая активность, класс "Z" - средняя активность.

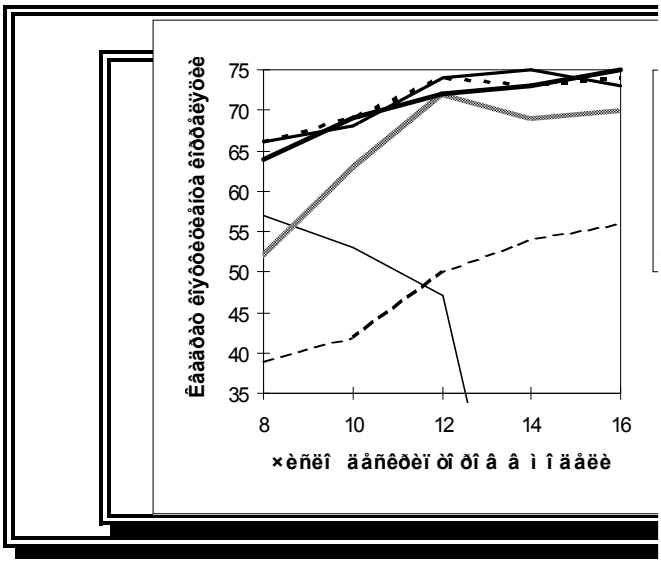


Рисунок 6.11 **Психотропная активность**. Зависимость R^2 - от числа дескрипторов, входящих в **QSAR** модель, для серий вычислений **I-VI**.

Рисунок 6.12 **Психотропная активность**. Зависимость R^2_{cv} - от числа дескрипторов, входящих в **QSAR** модель, для серий вычислений **I-VI** с использованием метода ***скользящего контроля***.

Была найдена система особых точек, расположенных на поверхности электростатического потенциала, которая не встречается в неактивных молекулах, но присутствует в активных структурах. Эта система представляет собой треугольник с параметрами: a,b=(**+215, +405**)Дж/моль, c=(-**545, -435**)Дж/моль; (ab)=(**9.3, 10.1**)A, (ac),(bc)=(**8.3, 8.9**)A.

Полученные **QSAR**-модели **VI** серии вычислений были использованы для прогнозирования психотропной активности экспериментально не исследованных соединений. Была использована выборка, состоящая из **21** соединения. Результаты прогноза были сопоставлены с данными экспериментальных измерений активности. Для всех 4-х экспериментально исследованных соединений результаты прогноза **полностью совпали** с данными эксперимента.

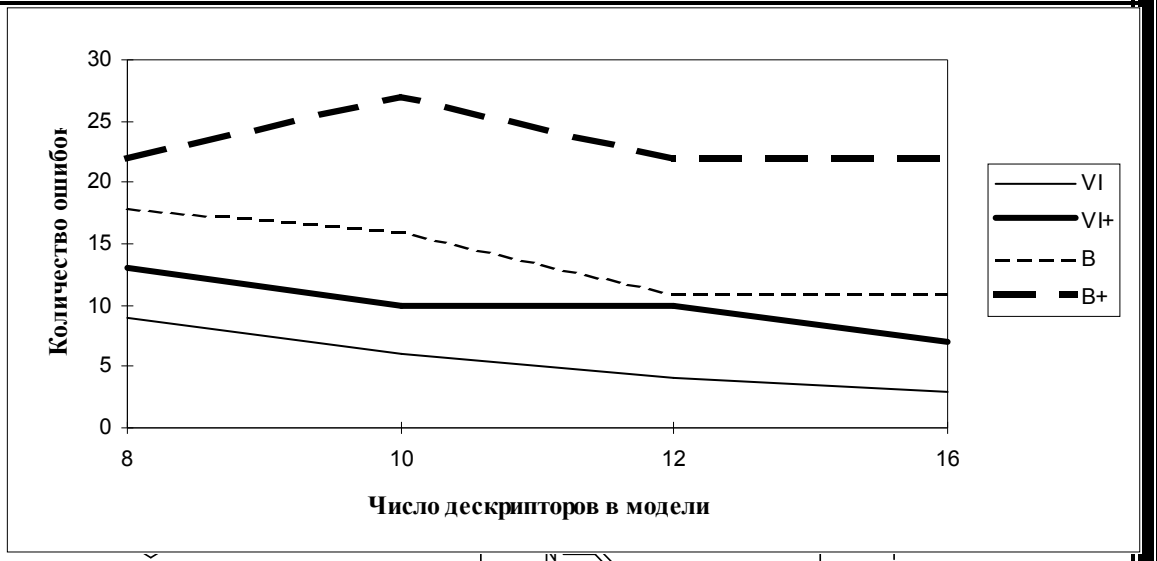


Рисунок 6.14 **Психотропная активность. Зависимость между количеством ошибок классификации и числом дескрипторов в модели. "VI" - кривая, построенная для серии VI, "B" кривая, построенная для** QSAR**-моделей системы БИБИГОН. "VI+" и "B+" -кривые для тех же моделей, но построенные с использованием метода *скользящего контроля*.**

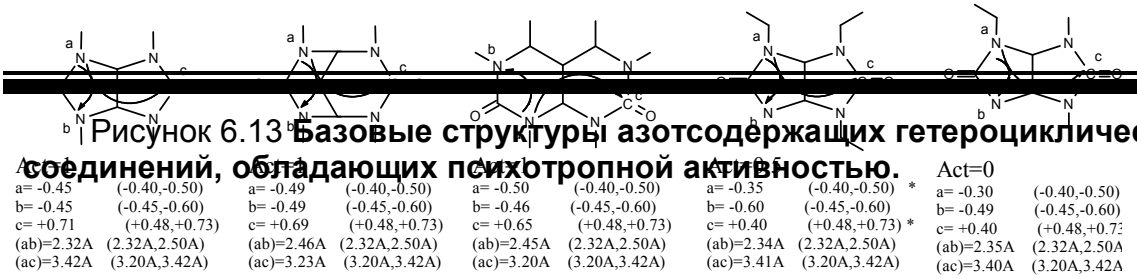


Рисунок 6.13 **Базовые структуры азотсодержащих гетероциклических соединений, обладающих психотропной активностью.**

Act=1		Act=1		Act=1		Act=0.5		Act=0	
a= -0.45	(-0.40,-0.50)	a= -0.49	(-0.40,-0.50)	a= -0.50	(-0.40,-0.50)	a= -0.35	(-0.40,-0.50) *	a= -0.30	(-0.40,-0.50)
b= -0.45	(-0.45,-0.60)	b= -0.49	(-0.45,-0.60)	b= -0.46	(-0.45,-0.60)	b= -0.60	(-0.45,-0.60)	b= -0.49	(-0.45,-0.60)
c= +0.71	(+0.48,+0.73)	c= +0.69	(+0.48,+0.73)	c= +0.65	(+0.48,+0.73)	c= +0.40	(+0.48,+0.73) *	c= +0.40	(+0.48,+0.73)
(ab)=2.32A	(2.32A,2.50A)	(ab)=2.46A	(2.32A,2.50A)	(ab)=2.45A	(2.32A,2.50A)	(ab)=2.34A	(2.32A,2.50A)	(ab)=2.35A	(2.32A,2.50A
(ac)=3.42A	(3.20A,3.42A)	(ac)=3.23A	(3.20A,3.42A)	(ac)=3.20A	(3.20A,3.42A)	(ac)=3.41A	(3.20A,3.42A)	(ac)=3.40A	(3.20A,3.42A

Рисунок 6.15 **Психотропная активность. Расположение систем особых точек, отвечающих за наличие психотропной активности.** Act=1 **- вещество обладает ярко выраженной активностью,** Act=0 **- вещество слабоактивно или неактивно вообще,** Act=0.5 **- вещество обладает промежуточной активностью;** a,b,c **- особые точки, расположенные на центрах атомов и входящие в состав систем особых точек; в качестве метки использован формальный заряд.** (ab),(ac) **- расстояния между особыми точками. В круглых скобках указаны допустимые пределы формального заряда и межточечных расстояний. * - данный параметр выходит за пределы, ** - указывает, что данный параметр сильно выходит за допустимые пределы.**

Анализируя результаты многочисленных вычислительных экспериментов, можно сформулировать **правило выбора особых точек:** для построения адекватного описания проявляемой веществами активности в качестве примитивов описания необходимо использовать точки, расположенные на поверхности электростатического потенциала, или же точки, расположенные на центрах атомов и маркированные величинам зарядов (возможно с использованием имени атома в качестве дополнительного маркера).

6.3 Выводы

Проведена практическая проверка метода и разработанных программ для построения **QSAR**-зависимостей различных свойств веществ: липофильность, температуры кипения и плавления, энтальпия образования, хроматографическая подвижность, поляризуемость, чувствительность к удару, антивирусная активность - на обучающих выборках органических соединений разных классов, включая гетероциклические соединения. Доказана прогностическая устойчивость найденных **QSAR**-зависимостей методом "скользящего контроля". Показано, что варьирование маркировки атомов и сложности структурных фрагментов существенно влияет на качество построенных **QSAR**-моделей. Рассмотрено применение метода к биологически активным соединениям: к триазинам с антибактериальной активностью, к соединениям с антигистаминной и галлюциногенной и анти-ВИЧ активностью, а также к ароматическим мускусным одорантам и азотсодержащим гетероциклическим веществам, обладающим психотропной активностью.

Приведены результаты ряда вычислительных экспериментов. Показано, что при одинаковом способе дискретизации непрерывных величин использование точек представления, расположенных на поверхности электростатического потенциала, приводит к результатам лучшим или сравнимым с результатами, получаемыми при использовании точек представления, связанных с центрами атомов. В то же время использование молекулярного электростатического потенциала приводит к повышению сложности модели и объема необходимых вычислений.

7. Заключение

Таким образом, в работе впервые разработана и успешно применена на практике методология решения задачи распознавания структурных объектов - молекулярных графов, - на основе поиска адекватных признаков (структурных дескрипторов), описывающих в ЭВМ особенности молекул при решении QSAR-задач. Основу методологии составляют:

- Многоуровневое представления молекулы как структурного объекта;
- Классификация локальных свойств атомов как примитивов описания молекул, которая определяет символьные маркеры атомов;
- Индуктивное порождение структурных спектров на основе выбранных примитивов молекул и автоматическое формирование таблицы "молекула -признак";
- Использование эволюционных алгоритмов - метода группового учета аргументов (МГУА):
    1. для анализа “очень широких” таблиц “молекула-признак”;
    2. для выбора описаний элементов структурного спектра, адекватно характеризующего свойство молекул.
- Проведение оценки свойств химических соединений на основе формирования базы знаний "структура-свойство", запросом к которой является молекулярный граф.

В работе проведена формализация основных этапов решения задачи "структура-свойство" и определена функциональная организация программной системы. Разработаны структуры данных для представления в ЭВМ ее основных объектов. Для селекции символьных структурных спектров, адекватных для исследуемого свойства, применены эволюционные алгоритмы. Работоспособность двух созданных QSAR-систем - BIBIGON и ChemAdd, - показана на практике. Важной отличительной чертой этих программ является их способность

анализировать "очень широкие" таблицы признаков, состоящие из нескольких тысяч столбцов.

В результате многочисленных вычислительных экспериментов построены QSAR-модели для прогнозирования свойств веществ различных химических классов. Достоверность и качество построенных зависимостей для прогнозирования свойств молекул показана с помощью метода "скользящего контроля". Характерной особенностью работы является то, что предложенная в ней методология может быть применена не только для анализа молекул, но и для построения моделей классификации *произвольных структурных объектов, задаваемых графами*, на основе генерации и селекции признаковых пространств, индуктивно порождаемых с использованием введенных примитивов.

Показано, что для биологической активности хорошие модели *могут быть* получены на топологическом уровне, что позволяет существенно упростить решение обратной QSAR-задачи. Заранее не известно, для какого биологического свойства можно отказаться от геометрической оптимизации структуры и отказаться от вычислений геометрических структурных дескрипторов.

Сформулированы общие принципы организации и функционирования QSAR системы для проведения полного цикла работ по прогнозированию свойств органических веществ на основе автоматического порождения и селекции структурных спектров молекул. Разработаны библиотеки объектов для создания QSAR систем, в которых реализованы алгоритмы анализа и символьной разметки молекулярных графов, алгоритмы автоматической генерации структурных спектров различной сложности, алгоритмы отбора значимых дескрипторов, алгоритмы формирования QSAR-моделей на основе МГУА, алгоритмы проверки прогностичности QSAR-моделей, алгоритмы оценки свойств на основе интерпретации QSAR-моделей и ряд других.

На основе созданных библиотек объектов разработаны две программных QSAR системы: BIBIGON и ChemAdd, которые позволяют автоматически строить аналоги аддитивных схем и решать “обратную QSAR-задачу” поиска

новых органических соединений, обладающих заданными свойствами путем комьютерной генерации и скрининга молекулярных графов.

В результате вычислительных экспериментов найдены признаковые пространства, в которых адекватно описываются молекулы, обладающие заданным свойством. На этой основе построены прогностичные QSAR-зависимости для ряда свойств веществ (как физико-химических, так и биологических) разных химических классов.

Предложена и реализована методика поиска и символьного представления знаний о QSAR-зависимостях.

В заключении пользуюсь случаем, чтобы выразить свою благодарность сотрудникам лаборатории математической химии ИОХ РАН и кафедры органической химии Химфака МГУ, совместная работа с которыми помогла преодолеть "химический барьер", необходимый для проведения формального анализа специфики QSAR-задач. Особая благодарность - аспирантам, работающим совместно с автором над задачей поиска "структура-свойство" и ее инструментальной поддержкой: Д.Ф.Митюшеву (ВМиК МГУ), И.Л.Зырянову (ИОХ РАН), Л.А.Пономаревой (ИОХ РАН), Г.М.Макееву (ИОХ РАН), Вороне Д.Г. (Мехмат МГУ), Чернобаеву А.А. (ВМиК МГУ), Юрченко Т.А. (ВМиК МГУ).

Настоящая работа не смогла бы появиться без финансовой поддержки проектов Российским фондом фундаментальных исследований (РФФИ).

Основные результаты диссертации опубликованы в следующих работах:

1. Кумсков М.И., Ломова О.А. Домашева Д.В. Лизунов А.К. *Алгоритм преобразования таблицы связности атомов в линейную нотацию Висвессера в диалоговом режиме* / Тезисы докл. 7-ой Всесоюзн. конф. "Использование вычислительных машин в спектроскопии молекул и химических исследованиях", Новосибирск, **1989**, с.173-174.

2. Кумсков М.И., Лизунов А.К. *Волновой алгоритм нахождения кольцевых структур молекулы по таблице связности атомов* / Тезисы докл. 7-ой Всесоюзн. конф. "Использование вычислительных машин в спектроскопии молекул и химических исследованиях", Новосибирск, **1989**, с.214-215.

3. Зефиров Н.С., Сапегин А.М., Палюлин В.А., Кумсков М.И., Раевский О.А. *"ИСТРА": интерактивная программная система*

компьютерного дизайна физиологически активных веществ / Тезисы докл. 7-ой Всесоюзн. конф. "Использование вычислительных машин в спектроскопии молекул и химических исследованиях", - Новосибирск, **1989**, с.228-229.

4. Кумсков М.И., Сухачев Д.В., Палюлин В.А., Ломова О.А. *Быстрая генерация М-графов на основе базовых структур в коде Висвессера* / Тезисы докл. межвуз. конф. "Молекулярные графы в химических исследованиях", Калинин, **1990**, с.54-55.

5. Кумсков М.И., Федоров А.Ю., Крехнов Б.В. *Программа предварительного отбора М-графов для прогнозирования их свойств.*/ Тезисы докл. межвуз. конф. "Молекулярные графы в химических исследованиях", - Калинин, **1990**, с.56-57.

6. Кумсков М.И., Пономарева Л.А., Зефиров Н.С*. Выбор алфавита структурных дескрипторов органических соединений при поиске зависимостей "структура-активность"* / Материалы 4-ой Всесоюзной школы-семинара "Статистический и дискретный анализ данных и экспертные оценки", - Одесса, **1991**, с.90-92.

7. Митюшев Д.Ф., Кумсков М.И., Зефиров Н.С. *Программа нахождения по молекулярному графу дескрипторов "тип химической структуры" в задаче "структура-активность"* / Материалы 4-ой Всесоюзной школы-семинара "Статистический и дискретный анализ данных и экспертные оценки", - Одесса, **1991**, с.315-317.

8. Пономарева Л.А., Кумсков М.И., Зефиров Н.С. *Формирование моделей ККСА на основе большого числа структурных дескрипторов. (Метод группового учета аргументов) /* Тезисы докл. 1-ой Всесоюзной конф. по теоретич. органической химии - ВАТОХ, Волгоград, **1991**, с.545.

9. Кумсков М.И., Пономарева В.А., Зефиров Н.С. "*Окраска" структурных дескрипторов органических соединений при поиске зависимостей "структура-активность" /* Тезисы докл. 1-ой Всесоюзной конф. по теоретич. органической химии - ВАТОХ, Волгоград, **1991**, с.550.

10.Зефиров М.И., Кумсков М.И., Пономарева Л.А., Митюшев Д.Ф., Смоленский Е.А. *Самоорганизация моделей - новое направление поиска зависимостей "структура-свойство*" / Материалы 9-ой Всесоюзной конф. "Химическая информатиия", - Черноголовка, **1992**, с.159-160.

11.Пономарева Л.А., Смоленский Е.А., Кумсков М.И., Митюшев Д.Ф., Зефиров Н.С. *Формирование структурных дескрипторов молекулярных графов*. / Материалы 9-ой Всесоюзной конф. "Химическая информатика", - Черноголовка, **1992**, с.96-97.

12.Пономарева Л.А., Олсуфьева Е.Н., Преображенская М.Н., Кумсков М.И., Зефиров Н.С. *Прогнозирование хроматографической подвижности антрациклиновых антибиотиков* / Тезисы докл. 9-ой Всесоюзной конф. "Химическая информатика", -Черноголовка, **1992**, с.98.

13.Пономарева Л.А., Олсуфьева Е.Н., Преображенская М.Н., Кумсков М.И., Зефиров Н.С. *Модель расчета хроматографической подвижности антрациклиновых антибиотиков ряда даунорубицина и его полусинтетических аналогов* / Хим.Фарм.Ж, **1993**, n.8, с.36-40.

14.Кумсков М.И., Пономарева Л.А., Смоленский Е.А., Митюшев Д.Ф. Зефиров Н.С. *Метод автоматического формирования структурных дескрипторов органических соединений* / Изв. РАН, Серия Химич., **1994**, т.8, с.1391-1394.

15.Кумсков М.И., Смоленский Е.А., Пономарева Л.А., Митюшев Д.Ф., Зефиров Н.С. *Системы структурных дескрипторов для решения задачи "структура-свойство* / ДАН, **1994**, т.336, n.1, с.64-66.

16.Митюшев Д.Ф., Кумсков М.И. *Алгоритмы обработки и анализа молекулярно-химических графов.* / Деп. в ВИНИТИ, ИОХ РАН, n. 288-В95 от 01.02.**1995**.

17.Кумсков М.И., Пономарева Л.А., Захарова М.В. *Новый подход к решению задачи QSAR для органических соединений* / Журн. Общей Химии, **1995**, т.65, вып.2, с.285-286.

18.Митюшев Д.Ф., Кумсков М.И. *Организация пользовательского интерфейса в программах поиска зависимостей "структура-свойство" химических соединений* / Деп. в ВИНИТИ, ИОХ РАН, n.289-В95 от 01.02.**1995**.

19.Кумсков М.И. *Перспективы использования программной системы BIBIGON для предсказания физико-химических свойств фторсодержащих органических соединений*. / Журн. Орг. Химии, 1995, т.31, Вып.10, с.1495-1498.

20.Зырянов И.Л., Кумсков М.И., Свитанько И.В. *Задача выбора "точек представления" молекулярного электростатического потенциала при поиске зависимостей "структура-биологическая активность* / Деп. в ВИНИТИ, ИОХ РАН, n. 2626-В-45 от 21.10.**1995**.

21.Макеев Г.М., Кумсков М.И. *Распознавание пространственных форм молекул биологически активных веществ* / Тезисы докл. 2-ой Всероссийской конф. "Распознавание образов и анализ изображений. Новые информационные технологии", - Ульяновск, **1995**, часть 3, с.152-154

22.Митюшев Д.Ф., Кумсков М.И., Петраускас А.А. *Комплексная программная система для прогнозирования свойств химических соединений на ПЭВМ* / Тезисы докл. 2-ой Всероссийской конф. "Распознавание образов и анализ изображений. Новые информационные технологии", - Ульяновск, **1995**, часть 4, с.136-138.

23.Макеев Г.М., Кумсков М.И*. Гибкая методика построения дескрипторов трехмерных структур* / Деп. в ВИНИТИ, ИОХ РАН, n.287-В95 от 01.02.**1995**.

24.Кумсков М.И., Митюшев Д.Ф., Петраускас А.А. *Формирование баз знаний "структура-свойство" химических веществ на основе анализа структурных спектров молекулярно-химических графов. соединений на ПЭВМ* / Тезисы докл. 2-ой Всероссийской конф. "Распознавание образов и анализ изображений. Новые информационные технологии", - Ульяновск, **1995**, часть 3, с.137-139.

25.Кумсков М.И., Митюшев Д.Ф. *Экспертное формирование метрик при распознавании молекулярно-химических графов. Непараметрическое*

оценивание свойств веществ / Тезисы докл. 2-ой Всероссийской конф. "Распознавание образов и анализ изображений. Новые информационные технологии", - Ульяновск, **1995**, часть 3, с.135-136.

26.Кумсков М.И., Митюшев Д.Ф. *Решение обратной задачи распознавания молекулярно-химических графов на основе построения минимальных путей в пространстве помеченных графов* / Тезисы докл. 2-ой Всероссийской конф. "Распознавание образов и анализ изображений. Новые информационные технологии", - Ульяновск, **1995**, часть 3, с.132-134.

27.Кумсков М.И., Митюшев Д.Ф., Пономарева Л.А., Пешкова С.Э. *Метод и программная система для автоматизированной экспресс оценки физико-химических и специальных свойств ВВ* / Материалы 21-ого Международного Семинара по Пиротехнике, М., **1995**, с.496-508.

28.Макеев Г.М., Кумсков М.И. *Распознавание пространственных форм молекул биологически активных веществ с целью классификации их свойств* / Тезисы докл. 7-ой Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.122-124.

29.Кумсков М.И., Зырянов И.Л., Макеев Г.М. *Учет локальных особенностей пространственных форм молекул с целью классификации их биологических свойств на основе структурных спектров* / Тезисы докл. 7-ой Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.116-117.

30.Пономарева Л.А., Митюшев Д.Ф., Кумсков М.И., Пешкова С.Э., Макеев Г.М. *Программа BIBIGON. Классификация свойств химических веществ на основе поиска обощенных структурных фрагментов молекул.* / Тезисы докл. 7-ой Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.142-143.

31.Пешкова С.Э., Кумсков М.И., Маслова Л.К. *Прогнозирование чувствительности к удару нитросоединений на основе структурных дескрипторов молекул* / Тезисы докл. 7-ой Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.137-138.

32.Пешкова С.Э., Кумсков М.И., Резчикова К.И. *Построение зависимостей "структура-свойство" для энергии активации высокоэнергетичных нитросоединений в конденсированной фазе с помощью структурных дескрипторов* / Тезисы докл. 7-й Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.138-140.

33.Апрышко Г.Н., Решетникова В.В, Лесная Н.А., Герасимова Г.К., Кумсков М.И., Маслова Л.К., Пономарева Л.А. *Прогнозирование цитотоксической активности химических соединений с помощью компьютерной системы BIBIGON* / Тезисы докл. 2-ого Национального конгресса "Человек и лекарство", - М., **1995**, с.57-58.

34.Кумсков М.И., Митюшев Д.Ф. *Применение метода группового учета аргументов для построения коллективных оценок свойств органических соединений на основе индуктивного перебора их*

“структурных спектров”. / Проблемы управления и информатики, **1996**, №4, с.127-149.

35.Кумсков М.И., Пешкова С.Э., Пономарева Л.А., Резчикова К.И. *Оценка энергии активации термического распада нитросоединений в газовой фазе на основе структурных дескрипторов* / Изв. РАН, Сер. Химич., **1996**, №8, с.1-4.

36.Lomova O.A., Sukhachev D.V., Kumskov M.I., Palyulin V.A., Tratch S.S., Zefirov N.S. *The generation of molecular graphs for QSAR studies by the acyclic fragment combining* / Comm. in Mathem. Chemistry (MATCH), **1992**, n.27, p.153-174.

37.Svitan'ko I.V., Kumskov M.I., Zyryanov I.L., Suslov I.A. *A method for describing the molecular electrostatic potential in determing structure-activity relationship* / Mendeleev Commun.,n.5, **1994**, p.161-162.

38.Kumskov M.I., Apryshko G.N., Gerasimova G.K., Maslova L.K *The prediction of the cytotoxic activity of fluorine containing compounds on personal computer* / Abstracts of the 1st Intern. Conf. "Chemistry and Application Fluorine Containing Compounds in the Industry" (CTAF'94), S.Peterburg, **1994**, c.116.

39.Kumskov M.I. *The use of BIBIGON system for prediction of properties of fluorine containing compounds on microcomputers* / Abstracts of the 1st Intern. Conf. "Chemistry and Application Fluorine Containing Compounds in the Industry" (CTAF'94), S.Peterburg, **1994**, c.114.

40.Apryshko G.N., Reshetnikova V.V., Gerasimova G.K., Kumskov M.I., Maslova L.K., Zefirov N.S. *The prediction of the cytotoxic activity of chemicals with the computerized BIBIGON system* / Abstacts of the 5th Intern. Symp. on Molecular Aspects of Chemotherapy. - Gdan'sk, Poland, **1995**, p.47.

41.Svitan'ko I.V., Kumskov M.I., Zyryanov I.L. *Molecular electrostatic potential method in determination of structure-activity relationship* / Abstracts of the 9th European Symp. on organic chemistry. - Warszawa, Poland, **1995**, p.256.

42.Kumskov M.I., Mityushev D.F., Oshchepkov E.N., Kolovanov E.A. *Computer-Aided Extraction and Application of Knowledge about Structure-Property Relationship in Factographical Chemical-Structure Data Bases* / Pattern Recognition and Image Analysis, **1995**, v.5, n.4, p.589-601.

43.Kumskov M.I., Zyryanov I.L., Svitan'ko I.V. *A New Method for Representing Spatial Electronic Structures of Molecules in the Problem of Structure-Biological Activity Relationship* / Pattern Recognition and Image Analysis, **1995**, v.5, n.3, p.477-484.

44.Kumskov M.I. *Information Models of Image Analysis Using “External World” as a Memory to Store the Objects under Study.*/ Pattern Recognition and Image Analysis, **1996**, v.6, n.1, p.199.

45.Svitan'ko I.V., Zyryanov I.L., Kumskov M.I., Khmel'nitskii L.I., Suvorova L.I., Kravchenko A.N., Markova T.B., Lebedev O.V., Orekhova G.A., Belova S.I *The surface molecular potential method for calculating the structure-activity relationship for psychotropic compounds* / Mendeleev Commun., **1995**, n.2, p.49-51

46.Kumskov M.I., Mityushev D.F. *Solving the Inverse Problem of Molecular Graph Recognition by Constructing Minimal Paths in a Labeled Graph Space* / Pattern Recognition and Image Analysis, **1996**, v.6, n.2, p.277-278.

47.Kumskov M.I., Mityushev D.F., Petrauskas A.A. *Generation of Structure-Property Chemical Substance Knowledge Bases from Analysis of Structural Spectra of Molecular Graphs* / Pattern Recognition and Image Analysis, **1996**, v.6, n.2, p.280-281.

48.Kumskov M.I., Mityushev D.F., Petrauskas A.A. *A Comprehensive Software System to Predict Properties of Chemical Compounds on a Personal Computer* / Pattern Recognition and Image Analysis, **1996**, v.6, n.2, p.412-413.

49.Makeev G.M., Kumskov M.I*. Recognition of Molecules of Biologically Active Substance* / Pattern Recognition and Image Analysis, **1996**, v.6, n.2, p.291-292.

50.Kumskov M.I., Mityushev D.F. *Expert Generation of Metrics in Recognition of Molecular Graphs: Nonparametric Estimation of Properties of Substances* / Pattern Recognition and Image Analysis, **1996**, v.6, n.2, p.279.

51.Makeev G.M., Kumskov M.I. *The Relationship between the Structure and Anti-AIDS Activity of Polyhydroxypiperidines and Polyhydroxypyrrolidines* / Mendeleev Commun., **1996,** n.1, p.27-29.

52.Yurchenko T.A., Kumskov M.I., Vorona D.G. *Metric Selection for Properties of Chemical Compounds Classification with the Help of Genetic Algorithm* / Extended Abstracts of the 4-th Open German-Russian Workshop "Pattern Recognition and Image Understanding", Valday, Russia, **1996**, p.153-156.

53.Chernobaev Al.A., Kumskov M.I., Mityushev D.F. *The Computer Generation of Marked Molecular Graphs Using Evilutionary Programming and Graph's Metrics* / Extended Abstracts of the 4-th Open German-Russian Workshop "Pattern Recognition and Image Understanding", Valday, Russia, **1996**, p.35-39.

54.Kumskov M.I., Sen'ko O.V., Peshkova S.E., Ponomareva L.A. *The Use of Two Stochastic Approximation Models for the Search of Quantitative Structure Activity Relationships (QSAR)* / Extended Abstracts of the 4-th Open German-Russian Workshop "Pattern Recognition and Image Understanding", Valday, Russia, **1996**,

55.Kumskov M.I., Chernobaev A.A., Mityushev D.F.*The Generation of Marked Molecular Graphs Using Evolutionary Programming and Graph's Metrics.*/ Proc. 1st Intern. Conf. on Evolutionary Computation and Its Application (EvCA'96), Moscow, **1996**, p.321-329

56.Kumskov M.I. *The QSAR Model Construction for Molecular Structure Classification on the Base of Feature's Spaces Searching and Selection* / Proc. 1st Intern. Conf. on Evolutionary Computation and Its Application (EvCA'96), Moscow, **1996**.

57.Kumskov M.I. *The construction quantitative "structure -activity" relationship (QSAR) by search of structural descriptor's spaces, which are adequate to the studied property*. / Proc. of QSAR Satellite Symp. of National

Congress of Pharmacologist of Ukraine (Poltava, 27-29 Sept. 1995), Poltava, **1996**, p.46-50.

58.Kumskov M.I., Zefirov N.S. *The program implementation of evolution method for building the property adapted index of organic compounds for the QSAR-task solution*. / Proc. of QSAR Satellite Symp. of National Congress of Pharmacologist of Ukraine (Poltava, 27-29 Sept. 1995), Poltava, **1996**, p.51-54.

59.Kumskov M.I., Mityushev D.F. *Group Method of Data Handling (GMDH) as Applied to Collective Property Estimation of Organic Compounds by an Inductive Search of Their Structural Spectra.* / Pattern Recognition and Image Analysis, **1996**, v.6, n.3, 497-509.

60.Makeev G.M., Kumskov M.I., Svitan'ko I.V., Zyryanov I.L. *Recognition of Spatial Molecular Shapes of Biologically Active Substances for Classification of Their Properties.*/ Pattern Recognition and Image Analysis, **1996**, v.6, n.4. p.795-808.

61.Mityushev D.F., Kumskov M.I. *An Integrated Software System to Predict Properties of Chemical Compounds on a Personal Computer.*/ Pattern Recognition and Image Analysis, **1996**, v.6, n.4. p.809-822.

62.Kumskov M.I., Chernobaev A.A., Mityushev D.F.*The Generation of Labeled Molecular Graphs Using Metrics and Evolutionary Programming.*/ Pattern Recognition and Image Analysis, **1997**, v.7, n.1. p.70-75.

8. Литература

1 Джурс О.П., Айзенауэр Т. Распознавание образов в химии: Пер.с англ., - М.: Мир, 1977, 248с.

2 Стьюпер Э., Брюггер У., Джурс П. *Машинный анализ связи химической структуры и биологической активности*: Пер.с англ./ М.: Мир, **1982**, 240с.

3 Кадыров Ч.Ш., Тюрина Л.А., Симонов В.Д., Семенов В.А. *Машинный поиск химических препаратов с заданными свойствами.* / Ташкент: Фан, 1989, 165с.

4 Журавлев Ю.И., Гуревич И.Б. *Распознавание образов и анализ изображений./* В сб. Искусственный интеллект. Справочник. т.2, - Под ред. Д.А.Поспелова, М.: Наука, 1990, с. 149-190.

5 Дьюар М. *Теория молекулярных орбиталей в органической химии*. М.: Мир, 1972, 590с.

6 Базилевский М.В. *Метод молекулярных орбит и реакционная способность органических молекул*. М.: Химия, **1969**, 303с.

7 Хедвиг П. *Прикладная квантовая химия*. М.: Мир, **1977**, 595с.

8 Бурштейн К.Я., Шорыгин П.П. *Квантовохимические расчеты в органической химии и молекулярной спектроскопии* М.: Наука, **1989**.-104с.

9 Кларк Т. *Компьютерная химия / Пер с англ.* М.: Мир, **1990**, с. 383

10 Ландау Л.Д., Лифшиц Е.М. *Квантовая механика*. Наука, **1974**, 752с.

11 Мак-Вини Р.,Сатклиф Б. *Квантовая механика молекул* М.: Мир, **1968**, 520с.

12 Пюльман Б., Пюльман А. *Квантовая биохимия*. М.: Мир, **1965**, 654с.

13 Волькенштейн М.В., Голованов И.Б., Соболев В.М. *Молекулярные орбитали в энзимологии.* М.: Наука, **1982**, 239с.

14 Kier L.B. *Molecular orbital theory in drug research*/ Acad. Press, New York, **1971**, 258p.

15 Слейтер Дж. *Методы самосогласованного поля для молекул и твердых тел*. М.: Мир, **1978**, 656с.

16 Дьячков П.Н. *Квантовохимические расчеты в изучении механизма действия и токсичности чужеродных веществ* / Итоги науки и техники ВИНИТИ, Серия Токсикология, **1990**, т.16, 280с.

17 Бушелев С.Н., Степанов Н.Ф. *Электронная структура и биологическая активность молекул.* М.: Знание, **1989**, 30с.

18 Richards W.G. *Quantum Pharmacology*. - Batterworths, London, **1983**.

19 Richards W.G. *Computer-Aided Molecular Design*. / Sci. Prog., Oxf., **1988**, v.72, p.481-492.

20 Dean P.M. *Molecular Recognition. The Measument and Search for Molecular Similarity in Ligand-Receptor Interaction*. / In: Johnson M.A. and Maggiora G.M. (Eds.) Concepts and application of molecular similarity. - New York, Willey, **1990**, p.211-238.

21 Bohm H-J. *A New Method for the De Novo Design of Emzyme Inhibitors. /* J. Comput.-Aided Mol. Design, **1992**, v.6, p.61-78.

22 Bohacek R.C., McMartin C. *Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth*. / J. Amer. Chem. Soc., **1994**, v.116, n.13, p.5560-5571.

23 Pearlman D.A., Murcko M.A. // J. Comput. Chem., **1993**, v.14, p.1184-1193.

24 Rotstein S.H., Murcko M.A. // J. Med. Chem., **1993**, v.36, p.1700-1710.

25 Moon J.B., Howe W.J. // Proteins: Struct. Funct. Genet., **1991**, v.11, p.314-328.

26 Nishibata Y., Itai A. *Atomic Creation of Drug Candidat Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation*. / Tetrahedron, **1991**, v.47, n.43, p.8985-8990.

27 Bondi A. *Van der Waals Volumes and Radii*. / J.Phys.Chem., **1964**, v.68, p.441-451.

28 Jorgensen W.L. *Rusting of the Lock and Key Model for Protein-Ligand Binding.* / Science, **1991**, n.254, p.854-855.

29 Wiley A.B., Rich D.L. *Peptidomimetics Derived from Natural Products*. / Med. Res. Rev., **1993**, v.13, p.327-384.

30 Martin Y. *Overview of Concepts and Methods in Computer-assisted Rational Drug Design.* / Methods Enzymol., **1991**, v.203, p.57-613.

31 Motoc I., Labanowski J.K., Bender C.F. *Molecular Design of Bioactive Molecules: Pharmacophore Model, Bioactive Conformation, and Three-Dimensional Quantitative Structure-Activity Relationship*. / MATH/CHEM/COMP 1988, Ed. by Graovac A. - Studies in Physical and Theor. Chem., v. 63, **1989**, p.235-244.

32 Jain A.N., Harris N.L., Park J. *Quantitative Binding Site Model Generation: Compass Applied to Multiple Chemotipes Tergeting the 5-HT1A Receptor* / J. Med. Chem., **1995**, v.38, p.1295-1308.

33 Marshall G.R., Motoc I. *Approaches to the Conformation of the Drug Bound to the Receptor.* / In: Molecular Graphics and Drug Design, Eds by Burgen A.S.V., Roberts G.C.K., Tute M.S., **1986**, Elsevier, Amsterdam.

34 Gund P. // Ann. Rep. Med. Chem., **1979**, v.14, p.299.

35 Bartmann W., Snatzke G (Eds) *Structure of Complexes between Bio-polymers and Low Molecular Weight Molecules*. - **1982**, Wiley-Heyden, Chichester.

36 Labanowski J.K, Motoc I., Naylor C.B., Mayer D., Dammkochler R.A.// Quant. Struct. Act. Relat., **1986**, v.5, p.138.

37 *SYBYL Molecular Modeling Package*, Tripos Associates, 1699 South Hanly Road, St. Louis, MO 63144, USA.

38 Mann A., Humblet C., Chambon J.P., Schlichter R., Desarmenien M., Feltz P, Wermuth C.G. // J. Med. Chem., **1985**, v.28, p.1440.

39 Hibert M.F., Gittos M.W., Middlemiss D.N., Mir A.K., Forzard J.R. // J. Med. Chem., **1988**, v.31, p.1087.

40 Doweyko A.M. *The Hypothetical Active Site Lattice. An Approach to Modelling Active Sites from Data on Inhibitor Molecules* / J. Med. Chem., **1988**, v.31, p.1396-1406.

41 Doweyko A.M. // J. Med. Chem., **1994**, v.37, p.1769-1778

42 Doweyko A.M., Mattes W.M. // Biochemistry., **1992**, v.31, p.9388-9392

43 Jain A.N., Koile K., Chapman D. *Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark.* / J. Med. Chem., **1994**, v.37, p.2315-2327.

44 Jain A.N., Dietterich T.G., Lathrop R.H., Chapman D., Critchlow R.E., Bauer B.E., Webster T.A., Lozano-Perz T. *Compass: A Shape-Based Machine Learning Tool for Drug Design.* / J. Comput. Aided Mol. Des., **1994**, v.8.

45 Dietterich T.G., Jain A.N., Lathrop R.H., Lozano-Perz T. *A Comparison of Dynamic Reposing and Tangent Distance for Drug Activity Prediction*. / In: Advances in Neural Information Processing System 6, Eds by Cowan J.D., Tesauro G., Alspector J., **1994**, Morgan Kaufmann, San Francisco, CA, p.216-223.

46 Goldberg D.E. *Genetic Algorithms in Search, Optimization and Machine Learning,* - Addison-Wesley, Wokingham, **1989**.

47 Davis L. (Ed) *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, **1991**.

48 Jones G., Willett P., Glen R.C. *A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation*. / J. Comput. Aided Mol. Des., **1995**, v.9, p.532-549.

49 Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., Ferrin T.E. // J. Mol. Biol., **1982**, v.161, p.269-288; Leach A.R., Kuntz I.D. // J. Comput. Chem., **1992**, v.13, p.730-748.

50 Lawerence M.C., Davis P.C. // Proteins: Struct. Funct., Genet., **1992**, v.12, p.31-41.

51 Martin Y.C., Bures M.G., Willett P. *Searching Databases of Three-Dimensional Structures* / In: Review in Computational Chemistry. Eds by Lipkowitz K., Boyd D. - **1990**, VCH Publishers, New York, p.213-263.

52 Martin Y.C. *Computer-Aided Design of Potentially Bioactive Molecules by Geometric Searching with ALLADIN* / Tetrahedron Comput. Methodol., **1990**, v.3, p.15-25.

53 Gund P., Wipke W.T., Langridge R. *Computer Searching of a Molecular Structure File for Pharmacophoric Patterns*. / Comput. Chem. Res., Educ., Technol., **1974**, v.3, p.5-21.

54 Gund P. *Three-Dimensional Pharmacophoric Pattern Searching* / Prog. Mol. Subcell. Biol., **1977**, v.11, p.117-143.

55 Van Drie J.H., Weininger D., Martin Y.C. *ALLADIN: An Integrated Tool for Computer-Assisted Molecular Design and Pharmacophore Recognition from Geometric,Steric and Substructure Searching of Three-Dimensional Molecular Structures.* / J. Comput.Aided Mol. Des., **1989**, v.3, p.225-251.

56 Jakes S.E., Willett P. *Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures. Selection of Inter-atomic Distance Screens. /* J. Mol. Graphics, **1986**, v.4, p.12-20.

57 Jakes S.E., Watts N., Willett P., Bawden D., Fisher J.D. *Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures. Evaluation of Search Performance*. / J. Mol. Graphics, **1987**, v.5, p.41-48;

58 Brint A.T., Willett P. *Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures. Comparison of Geometric Searching Algorithms.* / J. Mol. Graphics, **1987**, v.5, p.49-56.

59 Sheridan R.P., Venkateraghavan R. *Designing Novel Nicotinic Agonists by Searching a Database of Molecular Shapes*. / J. Comput. Aided. Mol. Des., **1987**, v.1, p.243-256.

60 Sheridan R.P., Rusinko A.III., Nilakantan R., Venkateraghavan R. *Searching for Pharmacophores in Large Coordinate Databases and Its Use in Drug Design*. / Proc. Natl. Acad. Sci., **1989**, v.86, p.8125-8169.

61 Murral N.W., Davies E.K. *Conformational Freedom in 3-D Databases*. / J. Chem. Inf. Comput. Sci., **1990**, v.30, p.312-316.

62 Burkert U., Allinger N.L. *Molecular Mechanics*. - Amer. Chem. Soc., Washington DC, 1982;

Буркет У., Эллинджер Н. *Молекулярная механика* / М.: Мир, **1989**.

63 Crippen G.M., Havel T.F. *Distance Geometry and Molecular Conformation*. - Research Studies Press, Wiley, New York, **1988**.

64 Rusinko A.III., Sheridan R.P., Nilakantan R., Haraki K. Bauman N., Venkateraghavan R. *Using CONCORD to Construct a Large Databases of Three-Dimensional Coordinate from Connection Table*. / J. Chem. Inf. Comput. Sci., **1989**, v.29, p.251-255.

65 Wipke W.T., Hahn M.A., *AIMB: Analogy and Intelligence in Model Building. System Description and Performance Cheracteristics*. / Tetrahedron Comp. Methodol., **1988**, v.1, p.141.

66 Dolata D.P., Leach A.R., Prout K. *WIZARD: AI in Conformation Analysis.* / J. Comput. Aided. Mol. Des., **1987**, v.1, p.73-85.;

67 Leach A.R., Prout K., Dolata D.P*. Automated Conformational Analysis: Algorithms for the Efficient Construction of Low-Energy Conformations*. / J. Comput. Aided. Mol. Des., **1990**, v.4, p.271-282.

68 Crippen G.M. *Review: Chemical Distance Geometry. Current Realization and Future Projection.* / J. Math. Chem., **1991**, v.6, p.307-324.

69 Martin Y.C. *3D Database Searching in Drug Design* / J. Med. Chem., **1992**, v.35, p.2145-2154.

70 Clark D.E., Willet P., Kenny P.W. *Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Implementation of Flexible Searching.* / J. Mol. Graphics, **1993**, v.11, p.146-156.

71 Clark D.E., Jones G., Willet P., Kenny P.W., Glen R.C. *Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Comformational-Searching Algorithms for Flexible Searching*. / J. Chem. Inf. Comput. Sci., **1994**, v.34., p.197-206.

72 Moock T.E., Henry D.R., Ozkabak A.G., Alamgir M. *Comformational Searching in ISIS/3D Databases.* / J. Chem. Inf. Comput. Sci., **1994**, v.34., p.184-189.

73 Milne G.W.A., Nicklaus M.C., Driscoll J.S., Wang Sh., Zahareviz D. *National Cancer Institute Drug Information System 3D Database*. / J. Chem. Inf. Comput. Sci., **1994**, v.34., p.1219-1224.

74 Фукунага К. *Введение в статистическую теорию распознавания образов:* Пер. с англ., - М.: Наука, **1979**, 368с.

75 Ту Дж., Гонсалес Р. *Принципы распознавания образов*: Пер.с англ. - М.: Мир, **1978**, 411с.

76 Martin Y.C. *Quantitative Drug Design: A Critical Introduction*, - Dekker, New York, **1978**.

77 Mager P.P *Multidimensional Pharmacochemistry: Design of Safer Drugs*, Academic Press, New York, **1984**; vol. 20.

78 Clementi, S. / In: *Drug Design: Fact or Fancy?,* Eds by Jolles G., Wooldridge K.R.H., - Academic Press, New York, **1984**, p.73-92.

79 Franke, R. *Theoretical Drug Design Methods*, - Elsevier, Amsterdam, **1984**, - Pharmacochemistry Library, vol.7.

80 Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. *Прикладная статистика: Классификация и снижение размерности*: Справ. изд. / Под ред. Айвазяна С.А. - М.: Финансы и статистика, **1989**, 607с.

81 Шараф М.А., Иламен Д.А., Ковальский Б.Р. *Хемометрика:* Пер. с англ. / Ленинград: Химия, **1989**, 269 с.

82 Hansch C, Fujita T. ro-sigma-pi-analysis*. A method for the correlation of biology activity and chemical structure* / J. Chem. Sci., **1964**, v.86, p.1616-1626.

83 Hansch C., Leo A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, - Wiley, New York, **1979**.

84 Free S.M., Wilson J.W. *A mathematical contribution to structure-activity studies* / J. Med. Chem. **1964**, v.7, n.4, p.395-399.

85 Хэнч К. *Об использовании количественных соотношений структура-активность (КССА) при конструировании лекарств* / Хим.-Фарм. Ж., **1980**, n.10, с.15-30.

86 Нижний С.В., Эпштейн Н.А. *Количественные- соотношения химическая структура - биологическая активность*. / Успехи химии, **1978**, №4, с.739-772.

87 Раевский О.А. *Введение в конструирование биологически активных веществ.* М., **1984**.

88 Раевский О.А., Сапегин А.М. *Развитие физико-химического подхода к распознаванию структур физиологически активных соединений*. / Хим. Фарм.Ж., **1987**, т.21, №11, с.1338-1341.

89 Сапегин А.М., Раевский О.А., Чистяков В.В., Мартынов И.В. *Алгоритм расчета донорно-акцепторных факторов для описания молекул биологически активных соединенийю* / Хим. Фарм.Ж., **1987**, т.21, №9, с.1098-1102.

90 Раевский О.А., Григорьев В.Ю. *Задачи по курсу "Конструирование биологически активных веществ*"., М., **1986**.

91 Сапегин А.М., Раздольский А.Н., Чистяков В.В., Раевский О.А. *РАСТР: Реализация физико-химического подхода к проблеме "структура-активность*". / Хим. Фарм.Ж., **1987**, т.21, №11, с.1341-1344.

92 Mekenyan O., Karabunarliev S., Bonchev D. *The Microcomputer OASIS System for Prediction the Biological Activity of Chemical Compounds*. / Compu. and Chem., **1990**, v.14, n.3, p.193-200.

93 Mekenyan O., Karabunarliev S., Bonchev D. *The OASIS Concept for Prediction the Biological Activity of Chemical Compounds*. / J. Math. Chem., **1990**, v.4, p.207-215.

94 Pastor M., Alverez-Builla J. *The EDISFAR Programs. Rational Drug Design*. / Quant. Struct. Act. Relat., **1991**, v.10, p.350-358.

95 Pastor M., Alverez-Builla J. *New Development ofEDISFAR Programs. Experimental Design in QSAR Practice*. / J. Chem. Inf. Comput. Sci., **1994**, v.34, p.570-575.

96 Kubinui H. *QSAR: Hansch Analysis and Related Approaches*, **1993**, VCH Verlag, Berlin,240p.

97 Waterbeemd H., Testa B. *The parametrization of Lipophilicity and Other Structural Properties in Drug Design.* / In: Advances in Drug Research, Ed. Testa B., **1987**, vol.16, Academic Press, New York, p.85-225.

98 Голендер В.Е., Розенблит А.Б. *Логико-комбинаторные методы конструирования лекарств* / Рига: Зинатне, **1983**, 302 с.

99 Waterbeemd H., Tayar N.E., Carrupt P.-A., Testa B. *Pattern recognition study of QSAR substituent descriptors*. / J. Comput. Aided Mol. Des., **1989**, v.3, p.111-132.

100 Green S.M., Marshall G.R. *3D-QSAR: a current perspective*. / Trends in Pharm. Sci., **1995**, v.16, p.285-291.

101 *Искусственный интеллект : применение в химии*. / Пер.с англ. ; Под ред. Пирса Т., Хони Б. - М.: Мир, **1988**, с. 410

102 Випке В.Т., Хан М.А. *Аналогии и рассуждения при построении моделей*. / В сб. Искусственный интеллект : применение в химии. / Пер.с англ. / Под ред. Пирса Т., Хони Б. - М.: Мир, **1988**, с. 410

103 Kubinyi H. (Ed*.) 3D-QSAR in Drug Design: Theory, Methods and Applications*. **1993**, ESCOM, Leiden.

104 Marshall G.R. // In 3D-QSAR in Drug Design: Theory, Methods and Applications. Kubinyi H. (Ed.) - **1993**, ESCOM, New York, p.80-116.

105 Internet WWW-site: http://199.217.144.2:8080/APPLICATION/GASP/doc/GASP.html

106 Cramer R.D.,III, Patterson D.E., Bunce J.D. *Comparative Molecular Field Analysis (CoMFA). 1.Effect of Shape on Binding of Steroids to Carrier Proteins*. / J.Amer.Chem. Soc., **1988**, v.110, n.18, p.5959-5967.

107 Dunn W.J.III, Wold S., Edlund U., Hellberg S. *Multivariate Structure-Activity Relationship between Data from Battery of Biological Tsts and Ensemble of Structure Descriptors. The PLS Method.* / Quant. Struct. Act. Relat., **1984**, v.3, p.131-137.

108 Hoskuldsson A. *PLS Regression Methods*. / J. Chemometrics, **1988**, v.2, p.211-228.

109 Харман Г. *Современный факторный анализ*: Пер. англ., - М.: Статистика, **1972**, 486с.

110 Иберла К. *Факторный анализ*: Пер. с англ., - М.: Статистика, **1980**, 389с.

111 *Алгоритмы и программы восстановления зависимостей*. / Под ред. Вапника, - М.: Наука, **1984**, 816с.

112 Cramer R.D.,III, Bunce J.D., Patterson D.E., Frank I.E. *Crossvalidation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies.* / Quant. Struct. Act. Relat., **1988**, v.7, p.18-25.

113 Cho S.J., Tropsha A. *Cross-Validated R2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achive Consistent Results. /* J.Med.Chem., **1995**, v.38, p.1060-1066.

114 Goodford P.J. *Drug Design by the Method of Receptor Fit.* / J. Med. Chem., 1984, v.27, p.557-564; Goodford P.J. // J. Med. Chem., **1985**, v.28, p.849-857.

115 Kim K.H., Martin Y.C. // J. Med. Chem., **1991**, v.34, p.2056-2060; Kim K.H., Martin Y.C. // Org. Chem., v.56, p.2723; Kim K.H., Martin Y.C. // Pharmacochem. Libr., v.16, p.151-154; Kim K.H. // Med. Chem. Res., v.2, p.22-27.

116 Davis A.M., Gensmantel N.P., Johansson E., Marriott D.P. // J. Med. Chem., **1994**, v.37, p.963-972.

117 DePriest S.A., Mayer D., Naylor C.B., Marshall G.R. // J. Am. Chem. Soc., **1993**, v.115, p.5372-5384.

118 Waller C.L., Marshall G.R. // J. Med. Chem., **1993**, v.36, p. 2390-2403.

119 Kellog G.E., Semus S.F., Abraham D. J. // J. Comput. Aided Mol. Des., **1991**, v.5, p.545-552.

120 Gaillard P., Carrupt P-A., Testa B., Boudon A. // J. Comput. Aided Mol. Des., **1994**, v.8, p.83-96.

121 Broto P., Moreau G., Vandycke C. *Molecular Structures: Perception Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor*. / Eur. J. Med. Chem., **1984**, v.19, n.1, p.66-70;

122 Broto P., Moreau G., Vandycke C. *Comparison of Logical or Three-Diminsional Molecular Structures by Means of Auto-Correlation Procedure*. / In: Computer Applications in Chemistry, Eds by Heller S.R., Potenzone R., Jr. - **1983**, Elsevier, Amsterdam, p.263-284.

123 Clark M., Cramer R.D.III, Jones D.M., Patterson D.E., Simeroth P.E. *Comparative Molecular Field Analysis (CoMFA). 2.Towards its use with 3-D Structural Databases*. / Tetrahedron Comput. Method.,**1990**, v.3, p.47-59.

124 Agarwal A., Pearson P.P., Taylor E.W., Li H.B. et al *Three-Dimensional Quantitative Structure-Activity Relationship of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: A Comparison of the Hansch and CoMFA Methods.* / J.Med.Chem., **1993**, v.36, p.4006-4014.

125 Cramer R.D.III, DePriest S.A., Patterson D.E., Hecht P., *The Developing Practice of Comparative Molecular Field Analysis*. / In: Kubinyi H. (Ed.) 3D-QSAR in Drug Design: Theory, Methods and Applications. **1993**, ESCOM, Leiden, p.443-458.

126 Tribaut U. *Application of CoMFA and Related 3D QSAR Approaches*. / In: Kubinyi H. (Ed.) 3D-QSAR in Drug Design: Theory, Methods and Applications. **1993**, ESCOM, Leiden, p.661-696.

127 Internet WWW-site: *http://www.tripos.com/acad/promot.html*

128 Baroni M., CostantinoG., Cruciani G., Riganelli D., Valigi R., Clementi S. *Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems*. / Quant. Struct. Act. Rel., **1993**, v.12, p.9-20.

129 Dunn III W.J., Wold S., Martin Y.C. *Structure-activity studies of betta-adrenergic agents using the SIMCA method of. pattern recognition*. / J.Med. Chem., **1978**, vol. 21, N9, p. 922-930.

130 Дюран Б., Оделл П. *Кластерный анализ*: Пер с англ., -М.: Статистика, **1977**, 128с.

131 Дуда Р., Харт К. *Распознование образов и анализ сцен*. / Пер. с англ. - М.: Мир, **1976**, 512с.

132 Горелик А.Л., Скрипкин В.А. *Методы распознавания*. / 2-е изд. - М., Высшая школа, **1984**, 208 с.

133 Jurs P.C., Isenhour T.L. *Chemical Applications of Pattern Recognition*, Wiley, New York, 1975; Джурс П., Айденауэр Т. *Распознавание образов в химии:* Пер.с англ. / М.: Мир, **1977**, 248 с.

134 Lipson H., Taylor C.A. *Fourier Transforms and X-Ray Diffractions*, G.Bell and Sons, London, **1958**.

135 Stuper A., Jurs P. *Classification of psychotropic drugs as sedatives or tranquilizers using pattern recognition techniques*. / J. Amer. Chem. Soc., **1975**, v. 97, n 1, P. 182-187.

136 King J.W., Kassel R.J., King B.B. *The Integrated Molecular Transform as a Correlation Parameter*. / Int. J. Quant. Chem: Quant. Biol. Symp., **1990**, v.17, p.27-34.

137 Famini G.R., Kassel R.J., King J.W., Wilson L.Y*. Using Theoretical Descriptors in Quantitative Structure-Activity Ralationship: Comparison with the Molecular Transform*. / Quant. Struct.-Act. Relat., **1991**, v.10, p.344-349.

138 Новиков В.П., Раевский О.А. *Представление молекулярной структуры в виде спектра межатомных расстояний для изучения связи структура- биологическая активностью* / Хим. Фарм.Ж., **1982**, т.16, №5, с.574-581.

139 Hopfinger A.J. *A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based upon Molecular Shape Analysis*. / J.Am. Chem. Soc., **1980**, v.102, p.7196-7206.

140 Rhyu K.-B., Patel H.C., Hopfinger A.J. *A 3D-QSAR Study of Anticoccidial Triazines Using Molecular Shape Analysis*. / J.Chem. Inf. Comp.Sci., **1995**, v.35, p.771-778.

141 Hofinger A.J., Burke B.J. *Molecular Shape Analysis. A Formalism to Quantitatively Establish Spatial Molecular Similarity*. / In: Johnson M.A. and Maggiora G.M. (Eds.) Concepts and application of molecular similarity. - New York, Willey, **1990**, p.173-209.

142 Moriguchi I., Jayoi K. *Use of van der Waals volume in structure-activity studies* / Chem. Bull., **1977**, v.25, n. 5. p.926-935.

143 Verloop A. *The STERIMOL Approach to Drug Design*. - Dekker, New York, **1978**.

144 Stanton D.T., Jurs P.C. *Development and Use of Charge Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies /* Anal.Chem., **1990**, v.62, p.2323-2329.

145 Stuper A.J., Brugger W.E., Jurs P.C. *Computer-Assisted Studies of Chemical Structure and Biological Function*, - Willey, New York, 1979;

146 Jurs P.C., Chou J.T., Yuan M., / In: Computer-Assisted Drug Design, Eds. Olson E.C., Chistoffersen R.E., - Amer. Chem. Soc., Washington, DC, **1979**, p.103-129.

147 Stuper A.J., Jurs P.C. *ADAPT: a computer system for automated data analysis using pattern recognition techniques*. / J. Chem. Inform. Comput. Sci., **1976**, v.16, n.2, p.99-110.

148 Bawden D. *Computerised chemical structure-handling techniques in structure-activity studies and molecular properties prediction*. / J. Chem. Inform. Comput. Sci., **1983**, v.23, n.1, p. 14-22.

149 Влэдуц Г.Э., Гейвандов Э.А. *Автоматизированные информационные системы для химии* / М.: Наука, **1974**, 310 с.

150 Bond V.L., Bowman C.M., Davidson L.C., Roush P.P., McGrew R.D., Williams D.G. *On-line storage and retrieval of chemical information. II.Substructure and biological-activity searching*. - J. Chem. Inform. Comput. Sci., **1979**, v.19, n.4, p.231-234.

151 Мищенко Г.Л., Гладкова Г.И. *О поиске групп структурно родственных соединений с помощью языка брутто-формул связей соединения*. - Научн. Техн.Инф., сер.2, **1980**, №12.

152 Cramer R. D., Redl G., Berkoff C. E. *Substructural Analysis. A Novel Approach to the Problem of Drug Design* / J. Med. Chem., **1974**, v. 17, p. 533-535.

153 Kowalski B.R., Bender C.F. *The application of pattern recognition to screening prospective anticancer drugs. Adenocarcinoma 755 biological activity test*. / J. Amer. Chem. Soc., **1974**, v.96, n.5, p. 916-918.

154 Adamson G., Bush J*. A comparison of the performance of some similarity and dissimilarity measures In the automatic classification of chemical structures*. / J. Chem. Inform. Comput. Sci., **1975**, v.15, n.1, p. 55-58.

155 Chu K.C, *Use of pattern recognition and cluster analysis to determine the pharmacological activity of some organic compounds*, / Anal. Chem., **1974**, v.46, n.9, p. 1181-1187.

156 Chu K.C., Feldman R.J., Shapiro M.B., Hazard G.F., Geran C.I*. Pattern recognition and structure-activity relationship studies. Computer-assisted prediction of antitumor activity in structurally diverse drugs in an experimental mouse brain tumor system*. / J. Med. Chem., **1975**, v.18, n.6, p. 539-545.

157 Hodes L. *Selection of molecular fragments for structure-activity studies in antitumor screening*. / J. Chem. Inform. Comput. Sci., **1981**, v. 21, n.3, p.132-136.

158 Cramer III R.D., Redl G., Berkoff C.E. Substructural analysis. A novel approach to the problem of drug design. / J. Med. Chem., **1974**, v.17, n.5, p. 533-535.

159 Hodes L., Hazard G.F., Geran R.I., Richman S*. A statistical heuristic method for automatic selection of drugs for screening*. / J. Med. Chem., **1977**, v. 20, n.4, p.469-475.

160 Adamson G.W., Bawden D. *A method of structure activity correlation using Wiswesser Line notation* / J. Chem. Inform. Comput. Sci., **1975**, v.15, n.4, p.215-220.

161 Adamson G.W., Bawden D. *Substructural analysis method of structure-activity correlation of heterocyclic compounds using Wiswesser Line notation* / J. Chem. Inform. Comput. Sci., **1977**, v.17, p.164-171.

162 Мищенко Г.Л. *Брутто-формулы связей соединений и их возможная роль при поиске фактографической информации в химии* / Информ. проблемы в современной химиию - М.: ВИНИТИ, **1976**, с.25-38.

163 Нигматуллин Р.С., Осипов А.Л., Пузаткин А.П., Коптюг В.А. *Статистический метод предсказания биологической активности многоатомных молекул на основе дескрипторов графов структурных формул.* / Хим. Фарм.Ж., **1985**, т.19, №2, с.179-185.

164 Brugger W.E., Stuper A.J., Jurs P.C. *Generation of descriptors from molecular structures*. / J.Chem. Inform. Comput. Sci., **1976**, v.16, n.2. p.105-110.

165 Jurs P.C., Chou J.T., Yuan M. *Computer-assisted studies of chemical carcinogens. A heterogeneous data set*. / J. Med. Chem., **1979**, v. 22, n.5,p.476-483.

166 Кабанкин А.С., Лексина Л. А., Бондарева А.П., Ландау М. А. *Анализ с помощью ЭВМ массива соединений, обладающих психотропной активностью. Применение топологических и подструктурных дескрипторов для распознавания вида психотропной активности* / Хим.-Фарм. Ж., **1985**, n.6, с.709-714.

167 Attias R., Dubois J.-E. *Substructure Systems: Concepts and Classifications* / J. Chem. Inf. Comp. Sci., **1990**, v.30, p.2-7.

168 Dubois J.-E. *DARC system in chemistry*. / In: Computer representation and manipulation of chemical information. Eds. by W.T.Wipke et al. - New York, Wiley-Interscience, **1974**, p. 239-264.

169 Attias R., *DARC.Substructure Search System: A New Approach to Chemical Information. /* J. Chem. Inf. Comp. Sci., **1983**, v.23, p.102-108.

170 Dubois J.-E., Panaye A., Attias R., *DARC System: Notion of Defined and Generic Structures. Filiation and Coding of FREL Substructure Classes*. / J. Chem. Inf. Comp. Sci., **1987**, v.27, p.74-82.

171 Авидон В. В. *Критерии сравнения химических структур и принципы построения информационного языка для информационно-логической системы по биологически активным соединениям* / Хим.-фарм. ж., **1974**. №. 8, С. 22-25.

172 Авидон В. В., Лексина А. А. *Дескрипторный язык для анализа сходства химических структур органических соединений* / Научн.-техн. информ., **1974**, Сер. 2, № 3, с.22-25.

173 Авидон В. В., Козлова С. П., Аролович В. С. *Кодирование циклических фрагментов на основе дескрипторного языка ФКСП* / Научн.-техн. информ., **1974**, Сер. 2., № 12, с.21-23.

174 Милль Дж.Ст., *Система логики силлогической и индуктивной. Изложение принципов доказательства в связи с методами научного исследования*. М., Изд. Лемана, **1914**. 880 с.

175 Забежало М.Н., Финн В.К., Авидон В.В., Катамадзе Т.Г., Блинова В.Г., Бодягин Д., Рабинков А.А. *Об экспериментах с базой данных с неполной информацией посредством ДСМ-метода порождения гипотез. /* НТИ, сер. 2, **1983**, № 3, с. 28-32.

176 Голендер В.Е., Розенблит А.Б. *Вычислительные методы конструирования лекарств* / Рига: Зинатне, **1978**, 238 с.

177 Забежайло М.Н, Финн В.К., Козлова С.П., Катамадзе Т.Г., Авидон В.В., Рабинков Д.А. *Об одном методе автоматического формирования гипотез и его программной реализации*. / НТИ, сер.2, **1982**, №4, с.20-26.

178 Финн В.К. *О машинно-ориентированной формализации правдоподобных рассуждении в стиле Ф. Бэкона - Д.С.Милля*. / Семиотика и информатика, **1983**, №20, с.35-101.

179 Гитлина Л. С., Голендер В. Е., Розенблит А. Б. *Метод отбора топологических и топографических признаков биологической активности химических соединений* / Хим.-фарм. ж., **1980**, Т.14, №.7, с.48-52.

180 Раевский О.А., Авидон В. В., Новиков В. П. *Использование унифицированной шкалы донорно-акцепторных взаимодействий для анализа сходства структур биологически активных соединений* / Хим.-Фарм. Ж., **1982**, n.8, с.968-971.

181 Franke R., Huebel S., Streich W.J. *Substructural QSAR Approaches and Topological Pharmacophores*. / Envir. Health Persp., **1985**, v.61, p.239-255.

182 Streich W.J., Franke R. *Topological Pharmacophores: New Methods and their Application. 1.The Methods LOGANA and LOCON*. / Quant. Struc. Act.Rel., **1985**, v.4, p.13-18.

183 Sheridan R., Venkataraghavan R. *New Methods in Computer-Aided Drug Design* / Acc. Chem. Res., **1987**, v.20, p.322-329

184 Carhart R.E., Smith D.H., Venkataraghavan R. *Atom pairs as molecular feature in structure-activity studies: Definition and applications*. / J. Chem. Inf. Comput. Sci., **1985**, v.25, p.64-73.

185 Nilakantan R., Bauman N., Dixon S.J., Venkataraghavan R. / J. Chem. Inf. Comput. Sci., **1987**, v.27, p.82-85.

186 Klopman G., Fercu D. *Application of the Multiple Computer Automated Structure Evaluation Methodology to a Quantitative Structure - Activity Relationship Study of Acidity* / Journal of Comput. Chemistry, **1994**, v.15, n.9, p.1041-1050.

187 Руврэ Д. *Следует ли заниматься разработкой топологических индексов? / В сб.: Химические приложения топологии и теории графов: Пер. с англ.*/Под ред. Р.Кинга.- М.: Мир, **1987**.с.183-205.

188 Татевский В.М. *Химическое строение углеводородов и закономерности в их физико-химических свойствах* / М.: Изд-во МГУ, **1953**, 320 с.

189 Татевский В.М., Бендерский В.А., Яровой С.С. *Методы расчета физико-химических свойств парафиновых углеводородов* / М.: Гостоптехиздат, **1960**, 114 с.

190 Татевский В.М., Бендерский В.А., Яровой С.С. *Методы расчета физико-химических свойств парафиновых углеводородов* / М.: Гостоптехиздат, **1960**, 114 с.

191 Benson S.W. *Thermochemical Kinetics*. 2nd ed., - Wiley: New York, 1976; Бенсон С. *Термохимическая кинетика* / М.: Мир, **1971**, 308 с.

192 Татевский В.М. *Классическая теория строения молекул и квантовая механика* / М.: Химия, **1973**, 516 с.

193 Яровой С.С*., Методы расчета физико-химических свойств углеводородов* / М.: Химия" 1**9**78, 256 с.

194 Татевский В.М. *Теория физико-химических свойств молекул и веществ* / М.: Изд-во МГУ, **1987**, 239 с.

195 Рид Р., Праусниц Дж., Шервуд Т. *Свойства газов и жидкостей* / Л.: Химия, 1982, 592 с.; Reid R.C., Prausnitz J.M., Sherwood T.K. *The properties of Gases and Liquids*, 3d ed., - McGraw-Hill, New York, **1977**.

196 Kier L.B., Hall L.H. *Molecular connectivity in chemistry and drug research* / Academic Press, New York, **1976**.

197 Lyman W.J., Reehl W.F., Rosenblatt D.H. (Eds) *Handbook of Chemical Property Estimation Methods. Environmental Behavior of Organic Compounds.* - McGrow-Hill, New York, **1982**.

198 Папулов Ю.Г., Татевский В.М. *Об учета попарных взаимодействий атомов, расположенных черех три атома в соединениях с тетраэдрической системой валентностей с формулой AnB2n+2* / Ж. Физ. Химии, **1962**, т.36, №1, с.189-206.

199 Папулов Ю.Г., Смоляков В.М. *Об эквивалентности феноменологических схем, построенных с учетом влияния ближайшего окружения атомов.* / Ж. Физ. Химии, **1972**, т.46, с.2933-2935.

200 Канович М.М., Папулов Ю.Г., Смоляков В.М. *О выборе параметров в аддитивных схемах расчета*. / Ж. Физ. Химии, **1983**, т.51, №3, с.748-751.

201 Smoljakov V.M., Papulov R.Yu. *Additivity schemes in atom-atomic approximation algorithm of comstruction and its application to substituted ethanes*. / Abstr. 7th American Conf. of Theor. Chem., **1990**, San Diego, p.128.

202 Смоляков В.М., Папулов Ю.Г., Поляков М.Н., Виноградова М.Г., Талызин И.В., Салтыкова М.Н., Ланцева О.В. *Использование теории графов для расчета термодинамических свойств углеводородов*. / Ж. Физ. Химии, **1996**, т.70, т.5, с.795-800.

203 Papulov Yu.G., Smoljakov V.M., Poljakov M.N. *On the use of topological indices in constructing additive calculation schemes.* / Abstr. IV Int. Conf. Math. and Comput. Chem., **1991**, Bled, Yugoslavia.

204 Папулов Ю.Г., Чернова Т.И., Смоляков В.М., Поляков М.Н. *Использование топологических индексов при построении корреляций "структура-свойство*". / Ж. Физ. Химии, **1993**, т.67, №2, с.203-209.

205 Смоляков В.М. *Зависимость свойств органических веществ от строения их молекул: расчетно-теоретическое исследование*. - Диссерт. на соиск. ученой степени доктора хим. наук в форме научного доклада, Тверь, Тверской Госуниверситет, **1995**.

206 Horvath A.L. *Molecular Design. Chemical Structure Generation from the Property of Pure Organic Compounds*. - Elsevier, Amstedam, **1992**. - Studies in Physical ans Theoretical Chemistry, v.75.

207 Sylvester J.J. *On an Application of the New Atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics*. / Amer. J. Math., **1878**, v.1, p.64.

208 Зыков А.А. Теория конечных графов. - Новосибирск: Наука. - Сибирское отделение, 1969.

209 Харари Ф. *Теория графов*. - М.: Мир, 1973.

210 Кристофидес Н. *Теория графов. Алгоритмический подход* / Перев. с англ. / М.: Мир, **1978**, с. 432.

211 Станкевич М.И., Станкевич И.В., Зефиров Н.С. *Топологические индексы в органической химии/ Успехи химии* - **1988**, т.57, N.3, с.337-336.

212 Basak S.C., Niemi G.J., Veith G.D. *Predicting Property of Molecules Using Graph Invariants*. / J. Math. Chem., **1991**, v.7, p.243-272.

213 Trinajstic N. *Chemical Graph Theory*, - v.1 and v.2, CRC Press, Boca Raton, FL, **1983**.

214 Kier L.B., Hall L.H. *Molecular Connectivity in Structure-Activity Analysis*, - Research Studies Press, Letchworth, UK,**1986**.

215 Kennedy J.W., Quintas L.V. *Application of Graphs in Chemistry and Physics*, - North-Holland, Amsterdam, **1988**.

216 Химические приложения топологии и теории графов: Пер с англ. / Под ред. Р.Кинга. - М.: Мир, **1987**, 560 с.

217 Руврэ Д. *Следует ли заниматься разработкой топологических индексов?* / В сб.: Химические приложения топологии и теории графов: Пер. с англ., под ред. Р.Кинга.- М.: Мир, **1987**, с.183-205.

218 Станкевич И.В. *Графы в структурной химии* / В сб.:Применение теории графов в химии. Под ред. Н.С.Зефирова, С.Н.Кучанова. - Новосибирск, Наука, **1988**, с.7-69.

219 Spialter L. / J. Amer. Chem. Soc., **1963**, v.85,p.2012.

220 Spialter L. / J. Chem. Doc., **1964**, v.4, p.269.

221 Schwenk A.J. / In: *New Direction in the Theory of Graphs*, Ed. Harary F., - Academic Press, New York, 1973.; Baker G.A. / J. Math. Phys., **1966**, v.7, p.2238.

222 Bonchev D., Trinajstic N. / J. Comput. Chem., **1981**, v.2, p.127.

223 Randic M. *On Molecular Identification Numbers* / J. Chem. Inf. Comp. Sci., **1984**, v.24, p.164-175.

224 Balaban A.T. *Highly Discriminating Distance-Based Topological Index*. / Chem. Phys. Lett., **1982**, v.80, p.399-404.

225 Niemi G.J., Regal R.R., Veith G.D. / In: Environmental Applications of Chemometrics, Eds. Breen J.J. and Robinson P.E. - Amer. Chem. Soc., DC, **1985**, p.148.

226 Basak S.C. / In: Practical Application of Quantitative Structure-Activity Relationship (QSAR) Study in Environmental Chemistry and Toxicology, Eds. Karcher W., DeViller J, - Kluwer Academic, New York, **1990**.

227 Константинова Е.В., Скоробогатов В.А. *Структурные и численные инварианты обыкновенных и молекулярных графов*. / Матем. методы в химич. информатике. - Новосибирск, **1991**, - Вычислит. системы, вып.140, с.87-129.

228 Скоробогатов В.А. *Относительные разбиения и слои графов*. / Вопросы обработки информации при проектировании систем. - Новосибирск, **1977**, - Вычислит. системы, вып. 69, с.3-10.

229 Mekenyan O., Bonchev D., Balaban A. *Topological indices for molecular fragments and new graph invariants*. / J. Math. Chem., **1988**, v.2, n.3, p.347-375.

230 Balaban A.T. *Confessions and reflections of graph-theoretical chemist*. / Comm. Math. Chem. (MATCH), **1993**, n.29, p.3-17.

231 Balaban A.T. *Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs*. / J. Chem. Inf. Comput. Sci., **1994**, v.34, p.398-402.

232 Balaban A.T., Balaban T.S. *New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances*. / J. Math. Chem., **1991**, v.8, p.383-397.

233 Balaban A.T. // J. Chem. Inf. Comp. Sci., **1992**, v.32, p.23.

234 Баскин И.И., СкворцоваМ.И., Станкевич И.В., Зефиров Н.С. *О базисе инвариантов помеченных молекулярных графов*. / Докл АН, **1994**, т.339, №3, с.346-350.

235 Wiener H. *Structural Determination of Paraffin Boiling Points*. / J. Am. Chem. Soc., **1947**, v.69, p.17-20.

236 Polansky O.E., Bonchev D. *The Wiener Number of Graphs*. / Comm. Math. Chem. (MATCH), **1986**, n.21, p.133-186.

237 Rouvray D.E. *The Role of the Topological Distance Matrix in Chemistry*. / In: Math. and Comput. Concepts in Chemistry, Ed. by Trinajstic N., - **1986**, Ellis Horwood, Chichester, England, p.295-306.

238 Tratch S.S., Stankevich M.I., Zefirov N.S. *Combinatorial models and algorithms in chemistry. The expanded Wiener number - A novel topological index.* / J. Comput. Chem., **1990**, v.11, p.899-908.

239 Randic M., Guo X., Oxley T., Krishnapriyan H., Naylor L. *Wiener Matrix Invariants*. / J. Chem. Inf. Comput. Sci., **1994**, v.34, p.361-367.

240 Randic M., Guo X., Oxley T., Krishnapriyan H. *Wiener Matrix: Source of novel graph invariants*. / J. Chem. Inf. Comput. Sci., **1993**, v.33, p.709-716.

241 Bogdanov B., Nikolic S., Trinajstic N. *On the Three-Dimensional Wiener Number.* / J. Math. Chem., **1989**, v.3, p.299-309.

242 Zhu H.-Y., Klein D.J., Lukovits I. *Extensions of the Wiener Number*. / J. Chem. Inf. Comput. Sci., **1996**, v.36, p.420-428.

243 Смоленский Е.А. *Применение теории графов к расчетам структурно-аддитивных свойств углеводородов* / Журн. Физ. Химии, **1964**, Т.38, N.5, с.1288-1290.

244 Скоробогатов В.А., Хворостов П.В., *Анализ метрических свойств графов / Методы обнаружения закономерностей с помощью ЭВМ*. - Новосибирск, **1981**, -: Вычислительные системы, вып.91, с.3-20.

245 Скоробогатов В.А. *Алгоритмический анализ молекулярных графов. (Основы метрического анализа*) / Новосибирск, НГУ, 1988, 84с.

246 Skorobogatov V.A., Dobrynin A.A. *Metric analysis of graphs* / Comm. Math. Chem. (MATCH), **1988**, n.23, p.105-151.

247 Balaban A.T. *Chemical Graphs. 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties.* / Comm. Math. Chem. (MATCH), **1986**, n.21, p.115-122.

248 Hosoya H. *Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Hydrocarbones*. / Bull. Chem. Soc. Japan, **1971**, v.44, p.2323-2339.

249 Hosoya, H. *Topological index as a sorting device for coding chemical structures*. / J. Chem. Doc., **1972**, v.12, n.3, p.181-183.

250 Trofimov M.I. *An Optimization of the Procedure for the Calculation of Hosoya's Index.* / J. Math. Chem., **1991**, v.8, p.327-332.

251 Hosoya H. *Topological Index as a Common Tool for Quantum Chemistry, Statistical Mechanics and Graph Theory*. / In: Math. and Comput. Concepts in Chemistry, Ed. by Trinajstic N., - **1986**, Ellis Horwood, Chichester, England, p.111-123.

252 Bonchev D., Trinajstic N. *Information theory, distance matrix, and molecular branching* / J. Chem. Phys., **1977**, v.67, n.10, p.4517-4533.

253 Bonchev D. *Information theoretic indexes for characterization of chemical structures*. Wiley, Chichester, **1983**.

254 Shannon C.E. // Bell Syst.Tech.J.,**1948**, v.27, p.379

255 Rouvray D.H. *The Evolution of the concept of Molecular Similarity*. / In: Johnson M.A. and Maggiora G.M. (Eds.) Concepts and application of molecular similarity. - New York, Willey, **1990**,

256 Johnson M.A. and Maggiora G.M. (Eds.) *Concepts and application of molecular similarity*. - New York, Willey, **1990**, 393p.

257 Jerman-Blazic B., Fabic-Petrac I., Randic M. *Evaluation of the molecular similarity and property prediction for QSAR purposes*. / Chemom. Intell. Lab. Syst., **1989**, vol.6, no.1, p.49-63.

258 Burt C., Richards W.G. *Molecular Similarity: The Introduction of Flexible Fitting*. / J. Comp. Aided Mol. Des., **1990**, vol.4, p.231-238.

259 Pepperrell C. *Three-Dimensional Chemical Similarity Searching*. - Research Studies Press, Taunton, Somerset, England, **1994**, 304p. (Computer and Chemical Information Series, vol.3)

260 Meyer A.M., Richards W.G. *Similarity of Molecular Shape*. / J. Comp. Aided Mol. Des., **1991**, vol.5, p.426-439.

261 Manaut M., Sanz F., Jose J., Milesi M. *Automatic Search for Maximum Similarity between Molecular Electrostatic Potential Distributions*. / J. Comp. Aided Mol. Des., **1991**, vol.5, p.371-380.

262 Drefahl A., Reinhard M. *Similarity-Based Search and Evaluation of Environmentally Relevant Properties for Organic Compounds in Combination with*

the Group Contribution Approach. / J. Chem. Inf. Comp. Sci., **1993**, vol.33, no.6, p.886- 95.

263 Randic M. *Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization*. / In: Johnson M.A. and Maggiora G.M. (Eds.) Concepts and application of molecular similarity. - New York, Willey, **1990**, p.77-145.

264 Журавлев Ю. И. *Об алгебраическом подходе к решению задач распознавания или классификации* / Проблемы кибернетики, М., Наука, 1978. Вып.33, с.5-68.

265 Гуревич И.Б. *Проблема распознавание изображений в задачах распознавания.* / В сб. Распознавание. Классификация. Прогноз. Математические методы и их применение. Вып.1, М.: Наука, **1989**, с.280-329.

266 Зыков А.А. *Теория конечных графов.* - Новосибирск: Наука, 1969.

267 Харари Ф. *Теория графов*. - Пер. с англ., - М.: Мир, **1973**, 300с.

268 Ивахненко А.Г., Зайченко Ю.П., Димитров В.Д. *Принятие решений на основе самоорганизации.* М.: Сов. радио, **1976**, 280с.

269 Ивахненко А.Г., Юрачковский Ю.П. *Моделирование сложных систем по экспериментальным данным.* М.: Радио и связь, **1987**, 120с

270 Шараф М.А., Шламен Д.А., Ковальский Б.Р. *Хемометрика / Пер. с англ.* - Ленинград: Химия, **1989**, 269с.

271 Журавлев Ю.И., Рязанов В.В., Богомолов В.П., Ворончихин В.А., Игнатенко О.Н., Ларин С.Б., Сенько О.В., Катериночкина Н.Н., Милютин С.В. *Параметрические логические модели анализа данных (программная система).*/ Тезисы докл. 7-ой Всероссийской Конф. "Математические методы распознавания образов", - Пущино, **1995**, с.95-96.

272 Вапник В.Н., Глазкова Т.Г., Кощеев В.А., Михальский А.И., Червоненкис А.Я. *Алгоритмы и программы восстановления зависимостей* / Под ред. Вапника В.Н. - М.: Наука, 1984, 816с.

273 Семенов Н.А. *Программы регрессионного анализа и прогнозирования временных рядов. Пакеты ПАРИС и МАВР.* - М.: Финансы и статистика, 1990, 111с.

274 Горелик А.Л., Гуревич И.Б., Скрипкин В.А. *Современное состояние проблемы распознавания*. *Некоторые аспекты.* - М.: Радио и связь, 1985, 162с.

275 Журавлев Ю.И. *Об алгебраическом подходе к решению задач распознавания или классификации /* Проблемы кибернетики, М.: Наука,1978, №33, с.5-68.

276 Журавлев Ю.И., Гуревич И.Б. *Распознавание образов и распознавание изображений.* / В сб. Распознавание. Классификация. Прогноз. Математические методы и их применение. Вып.2, М.: Наука, **1989**, с.5-72.

277 Журавлев Ю.И., Гуревич И.Б. *Распознавание образов и анализ изображений*. / Искусственный Интеллект. под ред. Попова Э.В., Книга 2, Справочник. - М.: Радио и связь, 1990, с.149-191.

278 Гренандер У. *Лекции по теории распознавания образов*. - Пер. с англ., - М.: Мир, 1979-1983, в 3-х томах.

279 Galil Z., Ukkonen E. (Eds) *Combinatorial Pattern Matching*. Proc. 6th Annual Symp., CPM 95, Espoo, Finland, July 5-7, 1995, - Berlin, Springer, **1995**. - Lecture Notes in Computer Science, v.937.

280 Горелик А.Л., Скрипкин В.А. *Методы распознавания.* - М.: Высшая школа, 1977, 117с.

281 Фукунага К. *Введение в статистическую теорию распознавания*. - М.: Наука, 1979, 368с.

282 Вапник В.А., Червоненкис А.Я. *Теория распознавания образов*. - М.: Наука, 1974, 415с.

283 Фу К.С. *Структурные методы в распознавании образов*. - Пер. с англ., - М.: Мир, 1977, 319с.

284 Журавлев Ю.И. *Об алгебраических методах в задачах распознавания и классификации. /* В сб. Распознавание. Классификация. Прогноз. Математические методы и их применение. - Вып.1, М.: Наука, **1988**, с.9-16.

285 Vapnik V.N. *The Nature of Statistical Learning Theory.* - New York, Springer-Verlag, 1995, 184p.

286 Айзерман М.А., Браверман Э.М., Розоноер Л.И. *Метод потенциальных функций в теории обучения машин* - М.: Наука, 1970, 240с.

287 Растригин Л.А. Эренштейн Р.Х. *Метод коллективного распознавания*. - М.? Энергоиздат, 1981, - Библиотека по автоматике, вып. 615, 80с.

288 Ледли Р.С. *Программирование и использовании цифровых вычислительных машин.* - Пер. с англ., М.: Мир, 1966, 644с.

289 Gourevitch I.B. *A Descriptive Method for Image Analysis Based on the Synthesis of an Image Model in the Class of Disjunctive Normal Forms.* / Pattern Recogn. Image Anal., v.5, n.3, p.356-363

290 Гуревич И.Б. *Предисловие к русскому изданию книги* Марр Д. Зрение. Информационный подход к изучению представления и обработки зрительных образов. - Пер. с англ., М.: Радио и связь, **1987**.

291 Марр Д. *Зрение. Информационный подход к изучению представления и обработки зрительных образов*. - Пер. с англ., М.: Радио и связь, **1987**, 400с.

292 Кохов В.А. *Метод количественного определения сходства графов на основе структурных спектров.* / Изв. РАН, Сер. Техническая. Кибернетика, **1994**, №5, с.143-159.

293 Кумсков М.И., Митюшев Д.Ф. *Применение метода группового учета аргументов (МГУА) для построения коллективных оценок свойств органических соединений на основе индуктивного перечисления их структурных спектров* / Проблемы управл. и информатики, **1996**, №4, с.127-149.

294 Kier L.B., Hall L.H. *Molecular connectivity in structure-activity analysis*. Willey: London, **1986**, 251p.

295 Кинг Р. *Химические приложения топологии и теории графов* / Пер. с англ. - М.: Мир. **1987**. 560с.

296 Скоробогатов В.А. *Алгоритмический анализ молекулярных графов*. Новосибирск, **1988**. 85 с.

297 Hansch C., Leo A.J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley: New York, **1979.**

298 Татевский В.М. *Химическое строение углеводородов и закономерности в их физико-химических свойствах*. М.: Изд-во МГУ, **1953**, 320 с.

299 Татевский В.М., Бендерский В.А., Яровой С.С. *Методы расчета физико-химических свойств парафиновых углеводородов*. М.: Гостоптехиздат, **1960**, 114с.

300 Кумсков М.И., Пономарева Л.А., Смоленский Е.А., Митюшев Д.Ф., Зефиров Н.С. *Метод автоматизированного формирования структурных дескрипторов органических соединений для количественных корреляций "структура-свойство"*. Изв. РАН, Серия химич., **1994**, N8, с. 1391-1393.

301 Кумсков М.И., Смоленский Е.А., Пономарева Л.А., Митюшев Д.Ф., Зефиров Н.С. *Системы структурных дескрипторов для решения задач "структура-свойство"*. ДАН, **1994**, т.336, N1, с.64-66.

302 Влэдуц Г.Э., Гейвандов З.А. *Автоматизированные информационные системы для химии.* - М., Наука, **1974**, 310с.

303 Smith E.G. *The Wiswesser Line-Formula Chemical Notation*. - McGraw-Hill, New York, **1968**, 310p.

304 Ишков А.Г., Кумсков М.И. *Линейная нотация Висвессера как компьютерная номенклатура химических соединений*. / Тезисы докл. 7-ой Всес. конф. "Использование вычислительных машин в спектроскопии молекул и химических исследованиях", Новосибирск, **1989**, с.168-169.

305 Марч Дж. *Органическая химия / Пер. с англ.* - М.: Мир, **1987**, т.1, 381с.

306 Пономарева Л.А., Смоленский Е.Н., Кумсков М.И., Митюшев Д.Ф., Зефиров Н.С. *Формирование структурных дескрипторов молекулярных графов* / Тезисы докл. IX Всес. конф. "Химическая информатика", Черноголовка, **1992**, с.96

307 Hall L.H., Mohney B., Kier L.B. *The electro topological state: structure information at the atomic level for molecular graphs.* / J.Chem. Inf. Comput. Sci., **1991**, v.31, p.76-82.

308 Balaban A.T. *Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs*. / J. Chem. Inf. Comput. Sci., **1994**, v.34, p.398-402

309 Берсукер П.Б., Димогло А.С., Горбачев М.Ю. *Электронно- топологический подход в исследовании связи структура-активность. Ингибирование тимидинфосфорилазы производными урацила*. / Биоорг. химия, т.13, №1, **1987**, с.38-44.

310 Лужков В.Б., Богданов Г.Н. *Квантовохимические расчеты в изучении противоопухолевых соединений*. / Успехи химии, **1986**, т.55, Вып.1, с.3-28.

311 Горбачев М.Ю., Димогло А.С., Берсукер И.Б., Влад П.Ф., Колца М.Н. *Роль электронных и структурных факторов в происхождении амбрового запаха у циклических соединений*. / Теорет. и эксп. химия, **1986**, №3, с.355-359.

312 Димогло А.С., Горбачев М.Ю., Берсукер И.Б. *Влияние структурных и электронных свойств производных урацила на ингибирование тримидин-фосфорилазы*. / Хим.-фарм. Ж, **1984**, №9, с.1086-1096.

313 Hall L.H., Mohney B., Kier L.B., *The electrotopological state: an atom index for QSAR.* / Quant. Struct.-Act. Relat., **1991**, v.10, p.43-51.

314 Basak S.C., Niemi G.J., Veith G.D. *Optimal Characterization of Structure for Predicting of Properties.* / J. Math. Chem., **1990**, v.4, p.185-205

315 Татевский В.М. Основы классической теории строения молекул. / М.: Изд-во Московского Университета, **1971**, 90с.

316 Ивахненко А.С., Зайченко Ю.П., Димитров В.Д. *Принятие решений на основе самоорганизации* / М.: Сов.Радио,1976, 280 с.

317 Ивахненко А.С., Юрачковский Ю.П. *Моделирование сложных систем по экспериментальным данным* / М.: Радио и связь,1987, 120 с.

318 *Statistical Methods for Digital Computers* / Eds by Enslein K., Ralson A., Wilf H. - N.Y.: John Wiley & Sons, Inc., **1977**, 386 p.

319 *Химические приложения топологии и теории графов./* Пер. с англ., Под ред. Р.Кинга. М.: Мир, **1987**. 560 с.

320 *Искусственный интеллект: применение в химии.:* Пер. с англ. / Под ред. Т.Пирса и Б.Хони. - М.: Мир, **1988**, 430 с.

321 Стьюпер Э., Брюггер У., Джурс П. *Машинный анализ связи химической структуры и биологической активности.:* пер, с англ. / Под ред. Евсеева А.М., М.: Мир, **1982**, 235 с.

322 Баскин И.И., Гордеева Е.В., Девдариани Р.О., Зефиров Н.С., Палюлин В.А., Станкевич М.И. *Методология решения обратной задачи в проблеме структура-свойство для некоторых топологических индексов. /* ДАН АН СССР, **1989**, т. 307, с. 613–617.

323 Алгоритмы и программы восстановления зависимостей / Под ред. Вапника В.Н., М.: Наука, **1984,** 816 с.

324 Кумсков М.И., Митюшев Д.Ф. Решение обратной задачи распознавания молекулярно-химических графов на основе построения минимальных

путей в пространстве помеченных графов / Тезисы докл. 2-ой Всероссийской конф. “Распознавание образов и анализ изображений. Новые информационные технологии”, Ульяновск, **1995**, часть 3, с.132-134.

325 Молодцов С.Г., Пиоттух-Пелецкий В.Н. *Конструирование неизоморфных графов по данному набору структурных параметров.* В сб. “Алгоритмы анализа структурной информации”, Вычислительные системы, Новосибирск, т.103, **1984**, с. 51.

326 Lomova O.A., Sukhachev D.V., Kumskov M.I., Palyulin V.A., Tratch S.S., Zefirov N.S. *The Generation of Molecular Graphs for QSAR Studies by the Acyclic fragment combining* / Commun. Mathem.Chem. (MATCH), **1992**, n.27, p.153-174.

327 Davis L. *Handbook of genetic algorithms* / Van Nonstrand Reinhold, New York, **1991**.

328 Chernobaev Al.A., Kumskov M.I., Mityushev D.F. *The Computer Generation of Marked Molecular Graphs Using Evolutionary Programming and Graph's Metrics* / Extended Abstracts of the 4-th Open German-Russian Workshop "Pattern Recognition and Image Understanding", Valday, Russia, **1996**, p.35-39.

329 Johnonson M., Naim M., Nicholson V., Tsai Ch.-C. *Unique Mathematical Features of the Substructure Metric Approach to Quantitative Molecular Similarity Analysis* / In: "Graph Theory and Topology in Chemistry", Eds. by R.B.King and D.H.Rouvray - Studies in Physical and Theoretical Chemistry, vol. 51, Elsevier., Amsterdam, **1987,** p.219-225.

330 Yurchenko T.A., Kumskov M.I., Vorona D.G. *Metric Selection for Properties of Chemical Compounds Classification with the Help of Genetic Algorithm. /* Extended Abstracts of the 4-th Open German-Russian Workshop "Pattern Recognition and Image Understanding", Valday, Russian Federation, **1996**.

331 Fogel D.B. *Applying evolutionary programming to selected control problems.* / Comput. Math. Applic., **1994**, vol.27, no.11,p.89-104.

332 Кумсков М.И., Митюшев Д.Ф. *Экспертное формирование метрик при распознавании молекулярно-химических графов. Непараметрическое оценивание свойств веществ* / Тезисы докладов 2-ой Всероссийской конференции,2-ой Всероссийской конф. "Распознавание образов и анализ

изображений. Новые информационные технологии", - Ульяновск, **1995**, стр.135-136.

333 Уотермен Д. *Руководство по экспертным системам / Пер. с англ.* - М.: Мир, **1989**, 388с.

334 Кумсков М.И., Митюшев Д.Ф., Ощепков Э.Н., Колованов Э.А.. *Автоматическое извлечение и использование знаний о зависимостях "структура-свойство" в фактографических базах структурных химических данных./* Pattern Recognition and Image Analisys, **1995**, Vol. 5, № 4, p. 567-579.

335 Баскин И.И., Палюлин В.А., Зефиров Н.С. *MODEL - программа интерактивного графического ввода молекулярных графов* /Тезисы докл. межвуз. конф. "Молекулярные графы в химических исследованиях" - Калинин, **1990**, с.6.

336 Кумсков М.И., Пономарева Л.А., Захарова М.В. *Новый подход к решению задачи QSAR для органических соединений* / Журн. Общей Химии, **1995**, т.65, вып.2, с.285-286.

337 Needham D.E., Wei I.C., Seybold R.G. *Experimental Values for the Physical Properties of the Alkans*. / J. Am. Chem. Soc.,**1988**, v.110, p.4186-4194.

338 DeTar F. // J.Org.Chem., **1991**, v.56, p.1478-1481.

339 Miller K.J. // J.Amer. Chem.Soc., **1990**, v.112, n.23, p.8533-8542.

340 Stanton D.T., Jurs P.C., Hicks M.G. // J. Chem. Inf. Comp. Sci., **1991**, v.31, n.2, p.301-323.

341 Camilleri P., Watts S.A., Boraston J.A. *A surface area approach to determination partial coefficients* / J. Chem. Soc. Perkin Trans. II, **1988**, N.9, p.1699-1707

342 Пономарева Л.А., Олсуфьева Е.Н., Преображенская М.Н., Кумсков М.И., Зефиров Н.С. *Модель расчета хроматографической подвижности антрациклиновых антибиотиков ряда даунорубицина и его полусинтетических аналогов* / Хим. Фарм. Ж, **1993**, n.8, с.36-40.

343 Bonchev D., Mekenyan O., Protic G., Trinajstic N. *Application of topological indexes to gas-chromatographic data: calculation of retention indexes of isomeric alkylbenzenes* / J. Chromatogr., **1979**, v.176, p.149-156.

344 Rekker R.F., Mannhold R (Eds) *Calculation of drug lipophilicity*. -/ Weinheim, VCH, **1992**, 115p.

345 Hansch C., Leo A.J. *Substituent Constants for Correlation Analysis in Chemistry and Biology,* - Wiley: New York, **1979.**

346 Марч Дж. *Органическая химия / Пер. с англ.* - М.: Мир, **1987**, т.1, 381с.

347 Leo A.J. *Calculating logP_{oct} from Structures*, / Chemical Reviews, **1993**, vol.*93*, p.1281.

348 Stothers J.B. *Carbon-13 NMR Spectroscopy*, Academic Press: New York and London, **1972**.

349 McFarland J.W., Cooper C.B., Newcomb D.M. *Linear Discriminant and Multiple Regression Analyses of Anticoccidial Triazines*. / J. Med. Chem., **1991**, v.34, p.1908-1911.

350 Зефиров Н.С., Петелин Д.Е., Палюлин В.А., Макфарланд Дж.У. // Докл.АН, 1992, т.327, №.4-6, с.504.

351 Fleet G.W.J., Karpas A., Dwek R.A. et al // *FEBS,**1988,***v.237; n.1,2;p.128-132

352 Смоленский Е.А., Гришина Г.В., Макеев Г.М., Зефиров Н.С.// Докл. Акад. Наук, **1993,** т.332, №5, с.603

353 DesJarlais R.L., Sheridan R.P., Dixon J.S., Kuntz I.D., Venkataraghan R., *Docking flexible ligands to macromolecular receptors by molecular shape.* / J.Med.Chem, 1986, v.29, p.2149-2153

354 Chothia J., Janin J. *Principles of protein-protein recognition.* / Nature, 1975, v.256, p.705-708. .

355 Wodak S.J., Janin J. *Computer analysis of protein-protein interaction*. / J.Mol.Biol., 1978, v.124, p.323-342.

356 Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., Ferrin T.E. *A geometric approach to macromolecule-ligand interactions.*/ J.Mol.Biol., 1982, v.161, p.268-269.

357 Hadkin E.E., Richards W.G. *Molecular Similarity Based on Electrostatic Potential and Electric Fields.*/ Int.J.Qant.Chem.:Qant.Biol.Symp., **1987**, v.14, p.105-110.

358 Sjoberg P. *The Use of the Electrostatic Potential at the Molecular Surface in Recognition Interractions: Dibenzo-p-dioxins and Related Systems.* / J. Mol.Graph., **1990**, v.8, p.81-84.

359 Richard A.M. *Quantitative Comparision of Molecular Electrostatic Potentials for Structure-Activity Studies.*// J.Comp.Chem., **1991**, v.12, p.959-969.

360 Crippen G.M. *Distance Geometry Approach to Rationalizing Binding Data* / J.Med.Chem., 1979, v.22, p.988-997.

361 Кумсков М.И., Зырянов И.Л., Свитанько И.В., *Новый метод представления пространственной электронной структуры молекул в задаче прогнозирования биологической активности* / Pattern Recognition and Image Analysis, 1995, n.3, p.477-484.

362 Colloe'h N., Mormon J. *A New Tool for Qualitative and Quantitative Analysis of Protein Surfaces Using B-Spline and Dencity of Surface Neighborhood.* // J.Mol.Graph., **1990**, v.8, p.133-140.

363 Svitan'ko I.V., Kumskov M.I., Zyryanov I.L., Suslov I.A. *A method for describing the molecular electrostatic potential in determing structure-activity relationship.* // Mendeleev Commun., **1994**, n.5, p.161-162.

364 Svitan'ko I.V., Zyryanov I.L., Kumskov M.I., Khmenitski L.I., Suvorova L.I., Kravchenko A.N., Markova T.B., Lebedev O.V., Orekhova G.A., Belova S.V. *The surface molecular potential method for calculating the structure-activity relathionship for psychotropic compounds.* / Mendeleev Commun., **1995**, n.2, p.49-51.

365 Буркерт У., Эллинджер Н.Л. *Молекулярная механика* // М.: Мир,**1986**.

366 Dewar M.J.S., Zoebisch E.G., Healy E.F., Stewart J.J.P. *AM1: Anew general purpose quantum mechanical molecular model* / J. Am. Chem. Soc., **1985**, v.107, p.3902-3909.

367 Stewart J.J.P. *MOPAC: a semiempirical molecular orbital program* / J. Comput. -Aid. Mol. Des., **1990**, v.4, p.90-105.

368 Димогло А.С. *Электронно-топологический подход в поиске и конструировании биологически активных химических соединений* / Автореферат дисертации на соискание ученой степ. докт. химич. наук, Ростов-на-Дону, **1987**, 48с.