547+541.69+615.015.11+681.3.06

Языки фрагментарного кодирования структуры соединений для компьютерного прогноза биологической активности

П. М. Васильев, А. А. Спасов

ПАВЕЛ МИХАЙЛОВИЧ ВАСИЛЬЕВ — кандидат биологических наук, старший научный сотрудник, докторант кафедры фармакологии Волгоградского государственного медицинского университета. Область научных интересов: QSAR, органическая химия, медицинская химия, фармакология. E-mail pmv@avtlg.ru

АЛЕКСАНДР АЛЕКСЕЕВИЧ СПАСОВ — член-корреспондент РАМН, доктор медицинских наук, профессор, проректор, заведующий кафедрой фармакологии Волгоградского государственного медицинского университета. Область научных интересов: фармакология, QSAR, биохимия, медицинская химия. E-mail farm@interdacom.ru

400131 Волгоград, пл. Павших борцов, д. 1, Волгоградский государственный медицинский университет, тел. (8442)97-15-34, факс (8442)38-30-28

Когда душа начинает понимать символ,
перед ней возникают представления,
недоступные чистому разуму.

Карл Густав Юнг

В 1861 г. на 36-м съезде немецких естествоиспытателей и врачей в Шпейере А.М. Бутлеров в своем докладе «О химическом строении вещества» утверждал, что «... химическая натура сложной частицы определяется натурой элементарных составных частей, количеством их и химическим строением»; о структурной формуле было сказано «... когда сделаются известными общие законы зависимости химических свойств тел от их химического строения, то подобная формула будет выражением всех этих свойств» [1]. Эти высказывания заключают в себе идею о важном значении химической структуры для познания способности веществ проявлять различные физико-химические и биологические свойства.

Современное понятие «химическая структура» весьма разноплановое и многоаспектное, однако графические плоские изображения молекул — их структурные формулы до сих пор остаются основным способом выражения информации о строении химических соединений. Именно эти «картинки» являются естественным языком химиков, именно с них начинается обсуждение тех или иных свойств конкретного вещества. По образному выражению академика Н.С. Зефирова, «структурная формула — это геном свойств химического соединения». Фактически это означает, что, имея в своем распоряжении адекватные способы параметризации двухмерной структурной формулы и методы извлечения содержащейся в ней информации, исследователь может получить, по нашим оценкам [2], до 90% сведений о свойствах изучаемого вещества.

Традиционно используемая и по сей день классическая фрагментация структурных формул соединений по функциональным группам, кратным связям, циклам, ароматическим или конденсированным системам [3] лежала в основе большинства ранних работ по исследованию соотношений «структура—биологическая активность». Она позволяла выявлять умозрительные эмпирические закономерности типа: «соединения, содержащие короткие ненасыщенные цепи, более активны, чем подобные им насыщенные соединения»; «введение алкильных радикалов в положения 1 или 3 уменьшает длительность действия соединений и наделяет их возбуждающим действием» [4].

Формирование во второй половине XX века научного направления QSAR как самостоятельного раздела науки потребовало разработки унифицированных способов кодирования структурных формул соединений совокупностью подструктурных фрагментов, удобных для использования в задачах вычислительного прогноза биологических и небиологических свойств веществ.

Неспециализированные информационно-поисковые языки

С появлением первых компьютеров и созданием больших химических информационно-поисковых систем возникла необходимость в разработке специальных информационных языков — *линейных нотаций*, позволяющих однозначно представлять двухмерную структурную формулу соединения набором строк символов, удобных для ввода в память ЭВМ с перфокарт. По основным принципам кодирования эти линейные химические номенклатуры являлись аналогами, наиболее известны из них линейная нотация Висвессера [5—7] и код Дайсона [8].

Любой язык, в том числе искусственный, имеет алфавит (систему неразложимых, уверенно отличимых друг от друга символов), словарь и грамматику, включающую правила словообразования, морфологию и

синтаксис [9]. Например, часть словаря линейной нотации Висвессера, кодирующая кислород, состоит из 4 букв и 3 морфов:

Q — гидроксил —OH

O — оксигруппа —O—

V — карбонил в кетонах =O

W — диоксогруппа, например, в $—NO_2$

VH — карбонил в альдегидах =O(H)

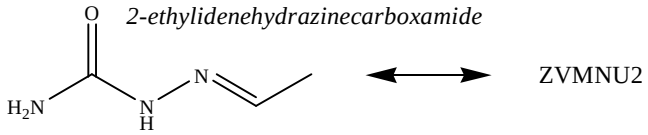
VQ — карбоксил —COOH

VO — сложноэфирная группа —CO—O—

В словаре нотации Висвессера буквенные обозначения присвоены всем химическим элементам (с учетом валентности их атомов и ближайшего окружения), а также основным типам связей, циклов и функциональных групп. Длина углеродной цепи кодируется числом, символ «&» разделяет коды цепей в точке их разветвления.

Грамматики первых линейных химических нотаций представляли собой варианты адаптированных правил номенклатуры органических соединений (старшинство групп, выбор главной цепи, начало ее нумерации и т.п.).

Пример кодирования по системе Висвессера структуры 2-этилиденгидразинкарбоксамида [7]:



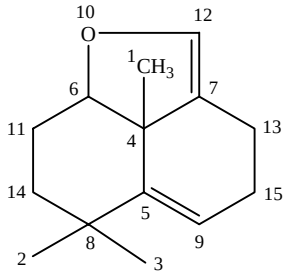
Здесь Z кодирует группу $—NH_2$, V — карбонил >C=O, M — группу —NH—, N — группу —N<, U обозначает двойную связь, 2 — длину углеродной цепи.

В исследовании [12] были рассчитаны регрессионные уравнения, связывающие реакционную способность соединений с наличием в их структуре различных подструктурных фрагментов, порождаемых из описания по линейной нотации химических структур из обучающей выборки.

Практически одновременно с линейными нотациями в химических информационно-поисковых системах для кодирования структуры соединений стали использоваться *матричные записи* — различные формы поатомной матрицы смежности, сначала в виде таблицы связей [13], затем в виде остовного дерева молекулярного графа (код Моргана) [14].

В таблице Моргана указываются порядковые номера и типы атомов, номер атома, с каким осуществляется связь при последовательном прохождении остовного дерева молекулярного графа, и тип этой связи (связь в цикле дополнительно обозначается символом «*»); в конце таблицы приводится перечень пар атомов, замыкающих циклы, с указанием типов связей.

Пример записи структуры производного гидронафтофурана в нотации Моргана (используется в CAS с 1965 г.) [10]:



Rank #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15							
Atom:	C	C	C	C	C	C	C	C	C	O	C	C	C	C	C							
Bond to:		8	8	1	4	4	4	5	5	6	6	7	7	11	13	10	12	8	14	9	15	
Bond is:		--	--	--	-*	-*	-*	-*	=*	-*	-*	=*	-*	-*	-*	RC	-*	RC	-*	RC	-*	

Следует отметить, что линейная нотация Висвессера и код Дайсона относятся к однозначным языкам, они позволяют полностью воспроизводить структурную формулу соединения по ее линейной записи, в связи с чем они использовались, например, в информационно-поисковой системе CAS — нотация Висвессера с 1953 г., код Дайсона с 1959 г. [10].

Фрагменты линейных кодов и их количественные характеристики ранее широко применялись в качестве подструктурных параметров для построения зависимостей «структура—свойство». В одной из первых таких работ [11] по выборке из 850 структурно разнородных соединений, проявляющих различные виды биологической активности, на основе описания по нотации Висвессера был сформирован первичный словарь признаков — подструктурных фрагментов. Затем для каждого вида активности методом подструктурного анализа были определены наиболее информативные признаки-фрагменты, которые в дальнейшем были использованы для конструирования новых активных соединений.

Код Моргана компактен, но весьма труден для восприятия неподготовленным пользователем. Поэтому интуитивно более понятный химику способ представления полного молекулярного графа в виде таблицы связности продолжает широко использоваться. Примером может служить текстовый коммуникативный формат «mol» фирмы MDL, который в настоящее время фактически является стандартом записи поатомной матрицы смежности [15].

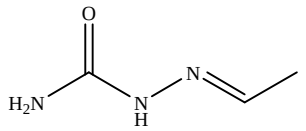
Описание структуры в mol-формате составляется из трех основных блоков:

1) строка параметров — содержит число атомов, число связей, число списков атомов, обозначение хиральности, версию формата;

2) список атомов — для каждого атома, помимо внутренних координат, указывается символ, отклонение от атомной массы основного изотопа, заряд, стереохимические особенности атома, число отдельно изображенных атомов водорода, нестандартная валентность;

3) список связей — для каждой пары связанных атомов указываются их номера, тип связи, ее стереохимические и топологические особенности.

Приведенная выше структура 2-этилиденгидразинкарбоксамида в mol-формате выглядит следующим образом:



```
  7  6  0  0  0  0  0  0  0  1  V2000
 −1,7860   −0,2062    0,0000   C  0  0  0  0  0  0  0  0  0  0  0  0
 −1,0716   −0,6187    0,0000   C  0  0  0  0  0  0  0  0  0  0  0  0
 −0,3572   −0,2062    0,0000   N  0  0  0  0  0  0  0  0  0  0  0  0
  0,3572   −0,6187    0,0000   N  0  0  0  0  0  0  0  0  0  0  0  0
  1,0716   −0,2062    0,0000   C  0  0  0  0  0  0  0  0  0  0  0  0
  1,7860   −0,6187    0,0000   N  0  0  0  0  0  0  0  0  0  0  0  0
  1,0716    0,6187    0,0000   O  0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0
  3  2  2  0  0  0
  4  3  1  0  0  0
  5  4  1  0  0  0
  6  5  1  0  0  0
  5  7  2  0  0  0
```

В строке параметров данной таблицы связности указано, что структура включает 7 атомов и 6 связей, она односвязанная, нехиральная, примечаний нет, версия формата 2000. Далее, например, первая строка списка атомов содержит xyz координаты атома углерода, он не заряжен, не асимметричен, не имеет отдельно присоединенных атомов водорода, стандартной валентности. В первой строке списка связей указано, что атомы 2 и 1 связаны одинарной связью без каких-либо особенностей.

Матричная форма удобна для однозначного представления структурных формул, но она не содержит в явном виде химически (и биологически) содержательной информации о структуре соединений — фактически это рабочий язык нижнего уровня, предназначенный для записи и хранения данных. Поэтому классические линейные нотации структуры, основанные на как можно более компактном индексировании химических символов буквами латинского алфавита и потому неудобочитаемые, в дальнейшем развивались параллельно с матричными — в сторону «химизации» и повышения наглядности представления структурной информации.

В настоящее время одной из наиболее развитых и популярных систем линейного кодирования структурных формул органических соединений является **система SMILES** — Simplified Molecular Input Line Entry System, упрощенная система ввода [структур] молекул [в форме] линейной записи [16—18].

SMILES имеет ряд достоинств, которые делают этот язык намного более понятным и удобным в обращении, чем предшествующие ему линейные нотации и матричные формы. Он представляет собой лингвистическую конструкцию, поэтому лучше воспринимается пользователем, чем более «компьютерная» таблица связей. Имеет простой и понятный словарь, основанный на обозначениях атомов и типов связей, привычных для химиков. Использует простую грамматику, состоящую из небольшого числа достаточно простых правил.

В основе грамматики SMILES лежат всего лишь шесть основных правил.

1. Атомы изображаются общепринятыми в химии символами.

2. Водородные атомы по умолчанию насыщают свободные валентности и не отображаются.

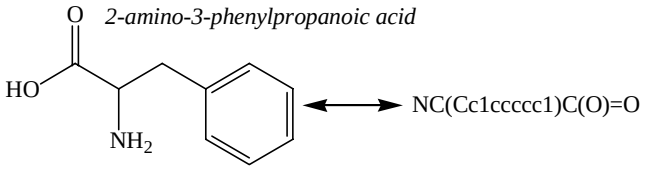
3. Соседствующие атомы записываются один за другим согласно порядка их следования в молекулярной цепи.

4. Двойная и тройная связи изображаются знаками «=» и «#», соответственно.

5. Коды боковых цепей заключаются в круглые скобки.

6. Циклы описываются путем присвоения цифровых индексов двум замыкающим цикл атомам.

В SMILES-нотации структура фенилаланина записывается так [19]:



Ниже приведены некоторые характерные примеры использования языка SMILES.

(1) В работе [20] описывается методика и программа расчета липофильности logP молекул на основе SMILES-кодов их структуры, с использованием которых генерируется расширенная матрица связей и оцениваются инкременты липофильности подструктурных атомных фрагментов. Итоговое значение липофильности получается в результате суммирования инкрементов подструктур, составляющих структуру исследуемого соединения.

(2) Разработанная в [21] расширенная версия языка SUPER-SMILES позволяет в больших химических базах данных легко находить пары-аналоги образа молекулы (в терминах SMILES-нотации), унифицировать и быстро выполнять процедуры модифицирования структур соединений: заменять атомы одних элементов на другие, протонировать молекулы, проверять валентность, расставлять атомы водорода (H-дополнение).

(3) В исследовании [22] выборка из 364 соединений, проявляющих различные виды биологической активности, была представлена множеством «миниотпечатков» («minifingerprints», MFPs) — коротких

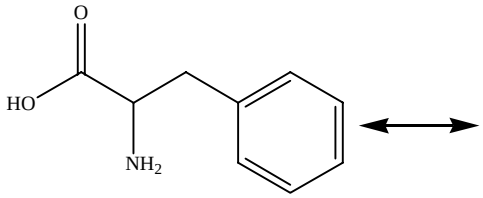
54-битных строковых кодов, сформированных по наличию или отсутствию в структуре молекулы подструктурных фрагментов, в свою очередь, формируемых на основе SMILES-описаний. Показано, что коды MFPs, записанные только на основе подструктурных фрагментов, позволяют производить распознавание биологически активных соединений методом сходства точнее, чем с помощью MFPs, сформированных на основе смешанных подструктурно-числовых параметров.

Система SMILES широко используется как текстовый коммуникативный формат при обмене структурными данными, поскольку она относится к однозначным языкам и SMILES-имя молекулы является синонимом ее структурной формулы.

В конце 1980-х годов для нужд номенклатуры Бейльштейна был разработан ROSDAL — язык, во многом похожий на SMILES, который, однако, не получил столь широкого распространения из-за неоднозначности, меньшей наглядности и более сложной системы грамматических правил [23].

Путем модификации SMILES фирма «Tripos» разработала **линейную нотацию SLN** (Sybyl Line Notation) — универсальный язык представления химической структуры, позволяющий кодировать «стандартные» органические соединения, полимеры и биомакромолекулы. Основное отличие SLN от SMILES — полная спецификация водородных атомов для всех «неводородных» атомов в структуре [24].

Для структуры фенилаланина SLN-запись выглядит следующим образом [19]:



C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH

SLN используется в QSAR исследованиях, для структурного и подструктурного поиска, в комбинаторных библиотеках.

Продолжает использоваться и традиционная функциональная фрагментация. Например, в системе Derwent World Patents Index [25, 26] разработан информационный **код CFC** (Chemical Fragmentation Codes), в котором описание химических соединений формируется в виде структур Маркуша и списков функциональных групп, ансамблей нескольких соседних атомов (прежде всего цепочек атомов), циклических систем и т.п.

Для фенилаланина это будут следующие фрагменты:

—OH >C=O —COOH —NH2 —Ph

Классическая фрагментация часто применяется для составления «отпечатков» (fingerprints) и «голограмм» структур соединений в форме так называемых счетчиков (counts), в которых указывается число различного типа атомов, связей или подструктурных фрагментов [19]. Обзор различных способов структурной фрагментации можно найти в [19, 27, 28].

Следует подчеркнуть, что варьируемая фрагментация без фиксированного словаря и грамматики в большинстве случаев носит субъективный характер и часто зависит от таких «нехимических» факторов, как доступность данных о структуре и свойствах соединений, способы предварительной обработки информации, методики построения решающих правил и т.п. Как следствие этого, словарь и грамматика языка изменяются при переходе от одной частной задачи к другой, что делает невозможным экстраполяцию полученных закономерностей в область неизученных соединений.

Специализированные языки для QSAR

Далее рассматриваются только достаточно полные системы линейного кодирования 2D-(двухмерных)-структурных формул органических соединений, имеющие детализированный (но не обязательно фиксированный) словарь фрагментных подструктурных дескрипторов и однозначно сформулированные грамматические правила их порождения. Такие информационные языки ориентированы на решение задач QSAR и входят как основные средства прогноза в состав самостоятельно существующих компьютерных программных комплексов.

Как это ни странно, в сравнении со структурной формулой специализированные языки фрагментарного кодирования должны быть вырожденными.

При любом фрагментарном кодировании структура соединения как единое целое «рассыпается». Часть информации при этом теряется, однако существует то преимущество, что исследователь получает возможность делать обобщения.

Подструктурная фрагментация позволяет формировать прогностические зависимости: наиболее часто встречающийся в структуре соединений фрагмент обладает более высокой экстраполирующей способностью, чем редкий фрагмент. Например, структуры большинства органических соединений содержат цепочку из двух последовательных σ-связей, соединяющих три «неводородных» атома, либо CH_3-группу; следовательно, закономерности, включающие в качестве параметров эти фрагменты, легко могут быть распространены на структуры, не присутствующие в обучающей выборке. Однако следует иметь ввиду, что при введении в прогнозные правила слишком большого числа таких параметров, резко снижается дискриминирующая способность этих правил, что приводит к увеличению ошибок классификации.

Таким образом, сбалансированный словарь подструктурного языка должен носить компромиссный характер: с одной стороны, оперировать достаточно общими для всех соединений фрагментами, позволяющими проводить экстраполяцию QSAR-зависимостей в область неисследованных соединений, с другой стороны, учитывать специфические подструктуры, позволяющие уверенно отличать соединения друг от друга.

По способу нотации подструктурные QSAR-языки можно условно разделить на два больших класса:

— основанные преимущественно на нотации вершин молекулярного графа (атомов, функциональных групп или циклов);

— основанные преимущественно на нотации ребер молекулярного графа (в простейшем случае типы связей).

Граница между этими классами весьма размыта, в большинстве существующих языков применяются оба подхода.

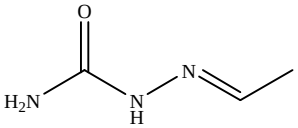
Простейшим способом формирования словаря подструктурного языка является использование для этой цели общеупотребимых фрагментов структуры: различного типа атомов, связей, функциональных групп, сопряженных, циклических, ароматических систем, а также их комбинаций. Семантика таких языков не включает в себя идей или концепций, отличных от общехимических. Такие языки (назовем их неконцептуальными) фактически представляют собой более или менее вырожденное описание структурной формулы соединения. Все эти языки неоднозначные.

Неконцептуальный язык, как правило, не имеет фиксированных словаря и грамматики. Они составляются в основном из функциональных групп в качестве подструктурных фрагментов и в зависимости от вида решаемых задач могут быть изменены путем введения новых типов фрагментов и правил их образования.

Рассмотрим наиболее характерные примеры неконцептуальных языков.

В системе CASE/MULTICASE, разработанной в 1984 г. [29, 30], структура соединений описывается цепочечными, разветвленными и циклическими подструктурными фрагментами, содержащими от 2 до 10 «неводородных» атомов, с указанием кратности связей.

Пример CASE-описания структуры 2-этилиденгидразинкарбоксамида:



CH3—CH	CH3—CH=N	CH3—CH=N—NH	CH3—CH=N—NH—C
CH=N	CH=N—NH	CH=N—NH—C	CH=N—NH—C—NH2
N—NH	N—NH—C	N—NH—C—NH2	CH=N—NH—C=O
C—NH	NH—C—NH2	N—NH—C=O	N—NH—C(=O)—NH2
C—NH2	NH—C=O	NH—C(=O)—NH2	
C=O	NH2—C=O		
CH3—CH=N—NH—C—NH2		CH3—CH=N—NH—C(=O)—NH2	
CH3—CH=N—NH—C=O			
CH=N—NH—C(=O)—NH2			

Классификационный алгоритм построен на основе биномиального распределения, используется также множественный регрессионный анализ [30]. В ходе анализа производится выделение «биофоров» и «биофобов» [31].

Как пример аналогичной CASE более простой нотации можно привести описание, основанное на одноатомных фрагментах (Atom/Fragment Contribution). Такое описание используется при расчете липофильности молекулы по аддитивной схеме [32].

Информационный язык, первоначально разработанный в 1987 г. для системы САПР-ПЕСТИЦИД [33], позднее стал использоваться в **системе CHANCE** [34, 35]. В этом языке для описания структуры составляется набор, в который входят следующие подструктуры:

— микрофрагменты;

— микрофрагменты с учетом первого окружения;

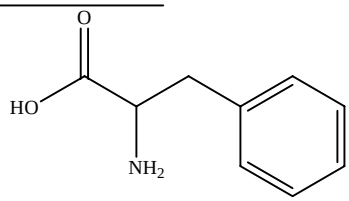
— микрофрагменты с учетом второго окружения;

— цепочки из двух и трех микрофрагментов с учетом типа связи.

При этом связи в ароматических циклах отдельно не кодируются.

В понятие «микрофрагмент» входят атомы с H-дополнением и меткой вхождения в цикл «*», функциональные группы, единичные и конденсированные алициклические группировки и гетероциклы.

Пример CHANCE-описания структуры фенилаланина с выделением в качестве микрофрагментов атомов с H-дополнением:

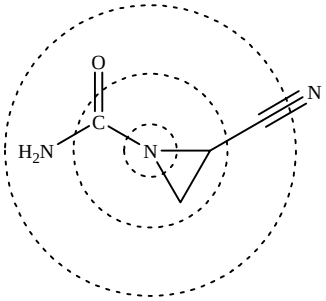


CH*	CH*—CH*	CH*—CH*=CH*	CH*(CH*,CH*)	CH*(CH*,CH*,CH*,CH*)
C*	CH*=CH*	CH*—CH*=C*	CH*(C*,CH*)	CH*(CH*,CH*,CH*,C*)
CH2	C*—CH*	CH*=CH*—C*	C*(CH*,CH*,CH2)	CH*(CH*,CH*,C*,CH*,CH2)
CH	C*=CH*	CH*=C*—CH2	CH2(C*,CH)	C*(CH*,CH*,CH*,CH*,CH2,CH)
C	C*—CH2	CH*—C*—CH2	CH(C,CH2,NH2)	CH2(C*,CH*,CH*,CH,C,NH2)
NH2	CH—CH2	C*—CH2—CH	NH2(CH)	CH(CH2,C*,NH2,C,O,OH)
O	C—CH	CH2—CH—NH2	C(CH,O,OH)	C(CH,CH2,NH2,O,OH)
OH	CH—NH2	CH2—CH—C	O(C)	O(C,CH,OH)
	C=O	CH—C=O	OH(C)	OH(C,CH,O)
	C—OH	CH—C—OH		

При классификации органических соединений на активные/неактивные используется алгоритм Байеса [33].

Примером похожей на CHANCE нотации является язык, основанный на описании химической структуры подструктурными фрагментами, центрированными на атом, связь или кольцо (аналогично ближайшему окружению микрофрагмента). Этот язык используется в системе классификации соединений методом сходства (Willett, 2000 г.) [36].

Рассмотрим кодирование атом-центрированными фрагментами структуры противоопухолевого препарата ВА 1 (2-циано-1-азиридинкарбоксамид):



Для каждого «неводородного» атома структуры анализируются «слои» связанных с ним других «неводородных» атомов, обычно до второго окружения. В приведенном примере «слой 0» содержит sp^3-гибридизованный атом азота (от которого строятся центрированные фрагменты); «слой 1» включает два sp^3- и один sp^2-гибридизованные атомы углерода; «слой 2» состоит из *sp*-атома углерода, sp^3-атома азота и sp^2-атома кислорода. Набор центрированных на этот атом фрагментов будет таким:

N_{sp3} C_{sp3},C_{sp3},C_{sp2} C_{sp},N_{sp3},O_{sp2}

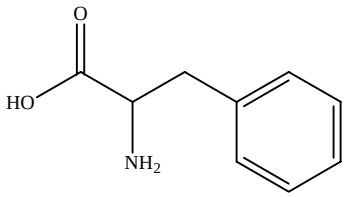
Полное атом-центрированное описание препарата ВА 1 выглядит следующим образом:

C_{sp3}	C_{sp3},N_{sp3}	C_{sp2},C_{sp}
C_{sp3}	C_{sp3},C_{sp},N_{sp3}	C_{sp2},N_{sp}
C_{sp2}	N_{sp3},N_{sp3},O_{sp2}	C_{sp3},C_{sp3}
C_{sp}	C_{sp3},N_{sp}	C_{sp3},N_{sp3}
N_{sp3}	C_{sp2}	N_{sp3},O_{sp2}
N_{sp3}	C_{sp3},C_{sp3},C_{sp2}	C_{sp},N_{sp3},O_{sp2}
N_{sp}	C_{sp}	C_{sp3}
O_{sp2}	C_{sp2}	N_{sp3},N_{sp3}

Различного типа центрированные фрагменты обычно используются для последующего формирования «отпечатков» («fingerprints») структур молекул.

В системе ОКАС, созданной в 1989 г. [37], более поздние модификации которой получили название **SARD** [38], основными элементами языка являются цепочки, состоящие из 1—3 простейших структурных фрагментов: $—CH_3$, $—CH_2—$, >CH—, >C<, кратные связи, функциональные группы, циклы.

Пример SARD-описания структуры фенилаланина:



Ph	(Ph)—(—CH2—)	(Ph)—(—CH2—)—(>CH—)
—CH2—	(—CH2—)—(>CH—)	(—CH2—)—(>CH—)—(—NH2)
>CH—	(>CH—)—(—NH2)	(—CH2—)—(>CH—)—(>C=O)
—NH2	(>CH—)—(>C=O)	(>CH—)—(>C=O)—(—OH)
>C=O	(>C=O)—(—OH)	(—NH2)—(>CH—)—(>C=O)
—OH		

В процессе предварительной обработки данных по первичному набору фрагментов структуры производится формирование сложных конъюнктивно-дизъюнктивных признаков с последующей статистической оценкой их информативности. Алгоритм прогноза основан на методе сходства к эталонам и методе простого голосования [37, 39].

В системе FALSE (1992 г.) для описания структуры органических соединений используются следующие подструктурные фрагменты [40, 41]:

— атомы, в том числе углерод с указанием типа гибридизации и водород с записью атома, к которому он присоединяется;

— типы связей, в том числе входящих в цикл;

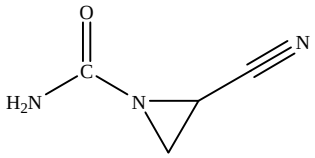
— функциональные группы (простейшие с учетом присоединенного к ним атома);

— цепочки длиной до 5 атомов, в том числе разветвленные;

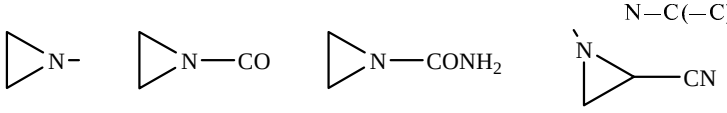
— единичные циклы с учетом гетероатомов и заместителей.

При прогнозе свойств органических соединений используется непараметрический дискриминантный анализ на основе теории нечетких множеств [40].

Пример FALSE-описания структуры противоопухолевого препарата ВА 1:



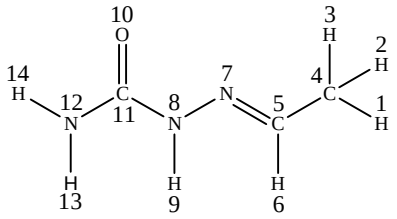
Csp^3	(C)—NH2	C—C	C—C—C	C—C—C—N	C—C—C—N—C
Csp^2	(C)—N<	C—N	C—C—N	C—C—C#N	C—C(—N)—C#N
Csp	(C)#N	C#N	C—C#N	C—C(—N)—C	C—C—N—C—N
H(—C)	(C)=O	C=O	C—N—C	C—C—N—C	C—C—N—C=O
H(—N)	>C=O		N—C—N	C—N(—C)—C	C—C—N(—C)—C
N	—CONH2		N—C=O	C—N—C—N	C—N—C—C#N
O	—CON<			C—N—C=O	C—N(—C)—C—N
=	—C#N			N—C—C#N	C—N(—C)—C=O
#				N—C(=O)—N	C—N—C(—N)=O
					C—N—C(=O)—N
					N—C—C—C#N
					N—C(—C)—C#N



Здесь =, # — двойная и тройная связь, соответственно.

Подструктурная часть языка **системы ILP** (1996 г.) использует в качестве слов названия химических элементов (атомов), составляющих соединения, и типы связей в соответствии с их квантовохимической классификацией [42, 43].

Пример подструктурной части ILP-описания 2-этилиденгидразинкарбоксамида (структуре присвоен условный порядковый номер 15):



atom(15, 15_1, hydrogen, _hydrogen, 0)	bond(15, 15_1, 15_4, single)
atom(15, 15_2, hydrogen, _hydrogen, 0)	bond(15, 15_2, 15_4, single)
atom(15, 15_3, hydrogen, _hydrogen, 0)	bond(15, 15_3, 15_4, single)
atom(15, 15_4, carbon, alkane_carbon, 0)	bond(15, 15_4, 15_5, single)
atom(15, 15_5, carbon, alkene_carbon, 0)	bond(15, 15_5, 15_6, single)
atom(15, 15_6, hydrogen, _hydrogen, 0)	bond(15, 15_5, 15_7, double)
atom(15, 15_7, nitrogen, imine_nitrogen, 0)	bond(15, 15_7, 15_8, single)
atom(15, 15_8, nitrogen, amide_nitrogen, 0)	bond(15, 15_8, 15_9, single)
atom(15, 15_9, hydrogen, amide_hydrogen, 0)	bond(15, 15_8, 15_11, single)
atom(15, 15_10, oxygen, carconyl_oxygen, 0)	bond(15, 15_10, 15_11, double)
atom(15, 15_11, carbon, carconyl_carbon, 0)	bond(15, 15_11, 15_12, single)
atom(15, 15_12, nitrogen, amide_nitrogen, 0)	bond(15, 15_12, 15_13, single)
atom(15, 15_13, hydrogen, amide_hydrogen, 0)	bond(15, 15_12, 15_14, single)
atom(15, 15_14, hydrogen, amide_hydrogen, 0)	

В ILP-описании для каждого атома указываются порядковый номер структуры в обучающей выборке, порядковый номер атома в структуре, тип атома, расчетное значение парциального заряда; для каждой связи указываются номер структуры, номера образующих связь атомов, тип связи.

Например, последняя строка из списка атомов в приведенном примере означает, что в структуре номер 15 атом под номером 14 является водородом, входящим в состав амидной группы, и для него парциальный заряд не рассчитывался; последняя строка из списка связей означает, что в структуре номер 15 атомы под номерами 12 и 14 соединены одинарной связью.

Для классификации система ILP использует логические решающие правила, формируемые с помощью языка программирования Progol [43].

Язык SubMat разработан в 2000 г. для кодирования структур в виде бинарных векторов с последующим формированием из них по всем соединениям исследуемой выборки обобщенной матрицы связности [44, 45]. Основными элементами словаря языка SubMat являются:

— метки атомов, в качестве которых выступают символы элементов, гетероатомы A, любые «неводородные» атомы Q;

— двухатомные фрагменты с записью типа связи (одинарная, двойная, тройная, связь в ароматическом цикле, прочая неспецифическая связь);

— одиночные неароматические циклы размером не более 8 атомов (из любых «неводородных» атомов, углеродные циклы, гетероциклы);

— неароматические бициклические системы;

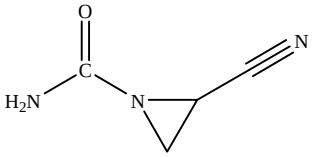
— ароматические моно- и бициклические системы с записью гетероатомов;

— прочие циклы;

— цепочки и разветвленные фрагменты (насыщенные углеводородные из 3—12 атомов, ненасыщенные углеводородные из 3—6 атомов с 1—2 кратными связями, все прочие из 3—4 атомов с хотя бы одним углеродным);

— функциональные группы.

Пример SubMat-описания структуры противоопухолевого препарата ВА 1:



C	C—C	C—C—C	C—C—C—N	Q—Q—Q—Q	—NH2	
N	C—N	C—C—N	C—C—C#N	Q—Q—Q#Q	—N<	
O	C#N	C—C#N	C—C(—N)—C	Q—Q(—Q)—Q	C#N	
A	C=O	C—N—C	C—C—N—C	Q—Q—Q=Q	>C=O	
Q		N—C—N	C—N(—C)—C	Q—Q(=Q)—Q	—CONH2	
		N—C=O	C—N—C—N		—CON<	
		Q—Q—Q	C—N—C=O			
		Q—Q=Q	N—C—C#N			
		Q—Q#Q	N—C(=O)—N			

Здесь # — тройная связь.

На основе языка SubMat реализованы процедуры распознавания биологически активных соединений методом сходства/различия [46].

Другой класс фрагментных подструктурных QSAR-языков составляют концептуальные (проблемно-ориентированные) языки. Базис таких языков составляют концепции, отражающие наиболее общие, но в то же время и наиболее характерные особенности проявления органическими соединениями тех или иных химических и/или биологических свойств. На основе базовых концепций формируются фиксированные словарь и грамматика, такие, чтобы их семантическое наполнение максимально соответствовало этим концепциям.

Слово или фраза концептуального языка — подструктурный фрагмент (дескриптор) несет в себе содержательную информацию о характере взаимодействия химических соединений с моделируемой системой.

Рассмотрим примеры концептуальных языков.

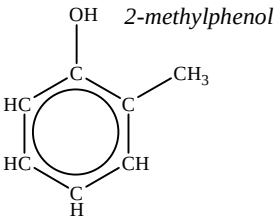
Язык FRAGMENT разработан в 1991—95 гг. специально для решения QSPR/QSAR задач [47, 48]. В основу этого языка положен ряд центральных положений теории реакционной способности органических соединений.

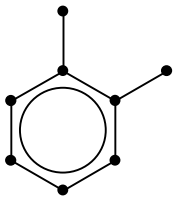
Формирование описания структуры соединения в этом языке состоит из двух этапов [47]. Первоначально производится генерация фрагментов на основе подробно классифицированных по окружению «неводородных» атомов с учетом их функциональности и типа связей. Затем ищутся все возможные обобщения классификации атомов во фрагментах каждого типа. Например, максимальным обобщением для атомов будет «•» — любой «неводородный» атом, одним из обобщений двух любых атомов будет «•—•» — фрагмент из двух «неводородных» атомов, связанных одинарной связью, и т.д. Таким образом, порождаются цепочки из 1—6 атомов, 3—6-членные циклы, несколько типов разветвленных фрагментов.

Алфавит и словарь языка FRAGMENT прост и понятен, в его основе лежит общепринятая химическая символика. Например, часть словаря, кодирующая основные элементы-органогены, выглядит так:

C	C_{sp3}	CR_2	CR_3	CHR	CHR_2	CHR_3	CH_2	CH_2R	CH_3	C_{A_r}	RC_{A_r}	HC_{A_r}
N	N_{sp3}	NR	NR_2	NH	NHR	NHR_2	NH_2					
O	OH	OR	S	SH	SR							
F	Cl	Br	I	Hal								

Приведем в качестве примера часть FRAGMENT-нотации структуры *о*-крезола с разным обобщением атомов (размер фрагментов не более 5 атомов).

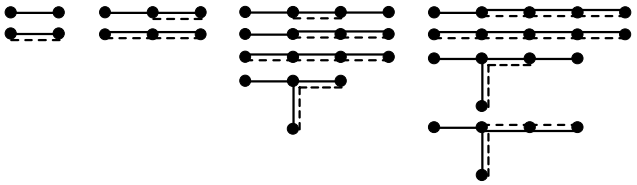




Обобщение на уровне функциональных групп:

CH_3	$CH_3—C_{A_r}$	$CH_3—C_{A_r}÷C_{A_r}$	$CH_3—C_{A_r}÷C_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷C_{A_r}÷HC_{A_r}÷HC_{A_r}$
C_{A_r}	$C_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷HC_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}$
HC_{A_r}	$C_{A_r}÷C_{A_r}$	$C_{A_r}÷C_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷C_{A_r}—OH$	$CH_3—C_{A_r}(÷HC_{A_r})÷C_{A_r}÷HC_{A_r}$
OH	$HC_{A_r}÷HC_{A_r}$	$C_{A_r}÷HC_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}(÷HC_{A_r})÷C_{A_r}$	$CH_3—C_{A_r}(÷C_{A_r})÷HC_{A_r}÷HC_{A_r}$
	$C_{A_r}—OH$	$C_{A_r}÷C_{A_r}—OH$	$C_{A_r}÷C_{A_r}÷HC_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}(÷HC_{A_r})÷C_{A_r}—OH$
		$HC_{A_r}÷HC_{A_r}÷HC_{A_r}$	$C_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}$	$CH_3—C_{A_r}÷C_{A_r}(—OH)÷HC_{A_r}$
		$HC_{A_r}÷C_{A_r}—OH$	$C_{A_r}÷C_{A_r}(—OH)÷HC_{A_r}$	$C_{A_r}÷C_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}$
			$HC_{A_r}÷HC_{A_r}÷C_{A_r}—OH$	$C_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}$
			$HC_{A_r}÷C_{A_r}÷C_{A_r}—OH$	$C_{A_r}÷C_{A_r}(—OH)÷HC_{A_r}÷HC_{A_r}$
			$HC_{A_r}÷C_{A_r}÷C_{A_r}÷HC_{A_r}$	$HC_{A_r}÷HC_{A_r}÷C_{A_r}÷C_{A_r}—OH$
			$HC_{A_r}÷HC_{A_r}÷HC_{A_r}÷HC_{A_r}$	$HC_{A_r}÷HC_{A_r}÷HC_{A_r}÷C_{A_r}—OH$
				$HC_{A_r}÷C_{A_r}÷C_{A_r}(—OH)÷HC_{A_r}$
				$HC_{A_r}÷C_{A_r}÷C_{A_r}÷HC_{A_r}÷HC_{A_r}$

Максимальное обобщение:



FRAGMENT дает избыточное описание, что позволяет при построении QSAR-зависимостей учитывать уровень сложности моделируемой системы.

Данный язык успешно используется в системах EMMA (регрессионный анализ) и NASAWIN (нейросетевое моделирование) для прогноза различных свойств органических соединений: хроматографических параметров, температуры кипения [49], энтальпии сублимации [50], магнитной восприимчивости [51] и т.д.

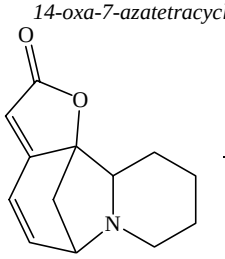
Язык SORTA разработан в 1996 г. как компонент системы SORT&gen для создания больших виртуальных химических библиотек [52]. В основу языка положена интерпретация правила Липински (правило отбора лекарственно-подобных соединений на основании некоторых свойств) [53], состоит он из дескрипторов циклов и гетероатомов.

Дескриптор циклов — это последовательность чисел, обозначающих размер всех циклов в структуре (размером не более 40 атомов), числа располагаются по убыванию размера циклов.

Дескриптор гетероатомов — это также набор чисел, отображающих все цепи между всеми парами гетероатомов $Q_1...Q_2$ в структуре, упорядоченные по значению «длина цепи +1»; последним всегда указывается код 01 самого гетероатома.

Пример SORTA-кодирования структуры алкалоида секуринина:

14-oxa-7-azatetracyclo[6.6.1.0^{1,11}.0^{2,7}]pentadeca-9,11-dien-13-one



14111009090706060505

15141411111008080807070504 01

(первый набор цифр — дескриптор циклов, второй — дескриптор гетероатомов).

Язык SORTA очень прост, что позволяет эффективно кластеризовать большие массивы структурной информации и отбирать в них биологически активные соединения по методу сходства.

Язык MNA (Multilevel Neighborhoods of Atoms, многоуровневые атомные окрестности) разработан в 1998 г. для системы прогноза спектра видов биологической активности PASS [54].

Дескрипторы MNA порождаются из таблицы связности в топологической аппроксимации — вырожденного представления молекулярного графа, в котором вершинами являются обобщенные метки атомов, а ребрами — единичные связи без указания их типа. Список обобщенных меток атомов [55] (см. табл. 1) сформирован на основе концепции биоизостеризма, согласно которой соединение, полученное путем замены в структуре молекулы атома одного химического элемента на атом другого элемента с идентичной внешней электронной оболочкой, должно обладать такими же биологическими свойствами, как и соединение-прототип [56].

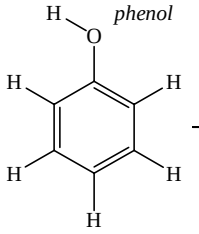
MNA-дескрипторы образуются следующим образом:

— дескриптор 0-го уровня представляет собой метку самого атома *A*;

— атом, находящийся в цепи (не в кольце) помечается префиксом «-»;

— дескриптор любого следующего уровня есть условное обозначение структурного фрагмента $A(D_1 D_2 \ldots D_i)$, где D_i является дескриптором предыдущего уровня для *i*-го непосредственного соседа данного атома с меткой *A*.

Пример MNA-кодирования структуры фенола [55]:



-H -H(-C) -H(C(CC-H))
-O -H(-O) -H(-O(C-H))
C -O(C-H) -O(C(CC-O)-H(-O))
C(CC-H) C(C(CC-H)C(CC-H)-H(C))
C(CC-O) C(C(CC-H)C(CC-H)-O(C-H))
C(C(CC-H)C(CC-O)-H(C))

В PASS для прогноза биологической активности используются MNA-дескрипторы первого и второго уровней [54, 55].

MNA является примером языка, основанного практически только на нотации вершин молекулярного графа, типы связей учитываются лишь косвенно, через число и тип атомов-соседей; это неоднозначный язык. Используется для прогноза видов биологической активности несколькими методами: вероятностно-статистическим [54], методом сходства [55], самосогласованной регрессии [57].

Система формирования голограмм структуры соединений — **язык HQSAR** — была специально разработана в 1997—2001 гг. для использования в модуле пакета программ Sybyl [58—60]. Структурная часть голограммы формируется на основе строк SLN-описания фрагментов [24].

Порождаются все подструктуры длиной от n до m атомов (значения n и m определяются из статистических соображений, обычно 4—7), содержащие цепочки, разветвления и циклы.

Таблица 1

Метки атомов языка MNA.

Метки с символом «*» объединяют несколько элементов. Особая метка R нотирует инертные газы, актиноиды и любой произвольный заместитель R

Метка	Элемент	Метка	Элемент	Метка	Элемент
H	H	Br	Br	Pd*	Pd, Pt, Au
C	C	Li*	Li, Na	Be*	Be, Zn, Cd, Hg
N	N	B*	B, Re	K*	K, Rb, Cs, Fr
O	O	Mg*	Mg, Mn	V*	V, Cr, Nb, Mo, Tc
F	F	Sn*	Sn, Pb	Ni*	Ni, Cu, Ge, Ru, Rh, Ag, Bi
Si	Si	Te*	Te, Po	In*	In, La, Ce, Pr, Nd, Pm, Sm, Eu
P	P	I*	I, At	Al*	Al, Ga, Y, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Tl
S	S	Os*	Os, Ir		
Cl	Cl	Sc*	Sc, Ti, Zr	R*	R, He, Ne, Ar, Kr, Xe, Rn, Ac, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr, Db, Jl
Ca	Ca	Fe*	Fe, Hf, Ta		
As	As	Co*	Co, Sb, W		
Se	Se	Sr*	Sr, Ba, Ra		

Параметры генерации задаются набором бинарных индексов (флагов):

— учета типа атома;

— учета типа связи;

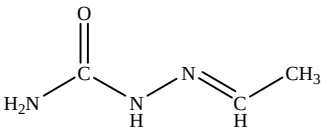
— учета числа связей атома;

— учета или игнорирования атомов водорода;

— учета хиральности атома;

— учета способности атома быть донором электронов, акцептором или бифильным.

Ниже приведен пример HQSAR-описания структуры 2-этилиденгидразинкарбоксамида с набором флагов генерации 111100 (1 — учитывать, 0 — нет):



CH3CH=NNH	CH3CH=NNHC	CH3CH=NNHCNH2	CH3CH=NNHC(NH2)=O
CH=NNHC	CH=NNHCNH2	CH3CH=NNHC=O	
NNHCNH2	CH=NNHC=O	CH=NNHC(NH2)=O	
NNHC=O	NNHC(NH2)=O		
NHC(NH2)=O			

Решающие правила в HQSAR рассчитываются на основе регрессионных моделей методом PLS [61].

Система HQSAR широко используется для прогноза самых разных видов биологической активности химических соединений; пример прогноза мутагенных свойств можно найти в работе [61].

Языки для QSAR семейства ФКСП

Особую группу концептуальных QSAR-языков составляют языки семейства ФКСП.

Созданный В.В. Авидоном в 1972 г. родоначальник этой группы **язык ФКСП** (фрагментарный код суперпозиции подструктур) [62—64], вероятно, был одним из первых языков, в котором органично соединились химическая и биологическая концепции. «Симбиоз» был настолько удачным, что на основе идей ФКСП впоследствии были разработаны несколько имеющих самостоятельное значение языков (МСДЦ, ФКСПм, QL). Некоторые идеи ФКСП легли в основу разработки языка MNA, хотя структура последнего очень отличается от структуры ФКСП.

Код ФКСП использовался как основной язык в нескольких системах прогноза биологической активности: в системе государственной регистрации химических соединений, для которой он и создавался [63, 65], в системе ORACL [66], в программах ДСМ-метода [67], в системе PASS [68].

Язык ФКСП основан на идеях теории химической рецепции [69], согласно которой биологический отклик живого организма на введение в него химического соединения формируется как результат специфического взаимодействия этого соединения со специализированным субмолекулярным комплексом (рецептором). Такое взаимодействие обеспечивается преимущественно за счет нековалентных взаимодействий — ионных, ион-дипольных, диполь-дипольных, дисперсионных, водородных и т.п. Для успешного связывания с рецептором в молекуле химического соединения должна существовать комплементарная активному центру рецептора группировка — фармакофор. При этом устойчивая фиксация молекулы на рецепторе возможна только в том случае, если в структуре фармакофора есть не менее двух активных центров связывания, называемых в ФКСП дескрипторными центрами (ДЦ). В качестве ДЦ могут выступать атомы или группы атомов с подвижной электронной системой — гетероатомы (для атомов основных элементов-органогенов учитываются типы связей и заряд), концевые атомы углерода, кратные и кумулированные двойные связи, ароматические системы в целом [62—64]. Перечень дескрипторных центров ФКСП и их кодов приведен в табл. 2.

Соответственно, потенциальные фармакофоры — дескрипторы языка ФКСП могут представлять собой либо два ДЦ, соединенные цепочкой углеродных атомов, либо циклические структуры, обеспечивающие многоцентровое взаимодействие.

На языке ФКСП структура соединения описывается совокупностью налагающихся друг на друга дескрипторов (отсюда название языка). Например, структура уксусной кислоты CH_3—C(=O)—OH может быть представлена тремя наложенными друг на друга фрагментами: CH_3—C—OH, CH_3—C=O и HO—C=O.

Каждый дескриптор имеет представление в виде числового кода.

В соответствии с этими постулатами словарь ФКСП включает два типа дескрипторов, линейных и циклических.

Линейные дескрипторы имеет структуру $ДЦ_1$—(длина цепи)—$ДЦ_2$—(индекс сопряжения). Длина цепи — это число атомов углерода в кратчайшем пути между двумя ДЦ, при этом путь не может проходить через другие ДЦ, кроме углеродсодержащих. Индекс сопряжения может принимать значения 1 (при наличии в цепи сопряжения) или 0 (сопряжения в цепи нет).

Особую форму линейных дескрипторов представляют собой дескрипторы замещения, в которых вместо длины цепи указывается код позиции замещения (число от 61 до 67) в зависимости от взаимного расположения ДЦ в цикле.

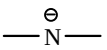
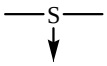
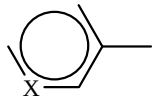
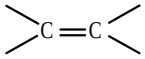
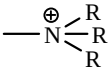
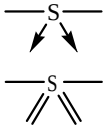
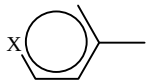
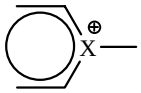
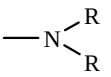
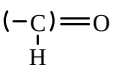
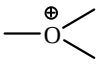
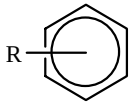
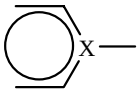
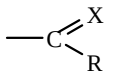
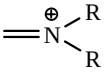
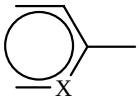
Любой линейный дескриптор имеет фиксированную длину и кодируется семизначным числом.

Циклические дескрипторы имеют структуру «голова»—«ядро»—«хвост». «Голова» кодирует геометрическую форму циклической системы: единичные циклы кодируются одной цифрой от 3 до 9, бициклические системы — двумя, полициклические — тремя и более цифрами и буквами, указывающими способ соединения циклов. «Ядро» указывает число π-электронов в циклической сопряженной системе (00, если сопряжения нет). «Хвост» кодирует типы и расположение гетероатомов. Каждый гетероатом обозначается специальным символом и номером его положения в циклической системе, таких кодов может быть несколько.

Таблица 2

Дескрипторные центры языка ФКСП.

Для каждого дескрипторного центра указан его числовой код и формула

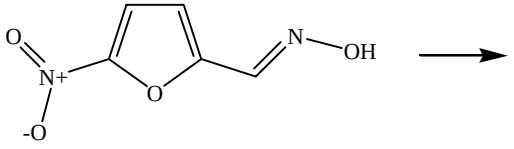
Код	ДЦ	Код	ДЦ	Код	ДЦ	Код	ДЦ	Код	ДЦ
00		11	—OH	23		36		46	
01		12	—OR	24		37		47	P
02	—NH_2 —NHR	13	=O (кроме альдегидного)	25	=S	40		51	As Sb
03		14	 (альдегидный)	31	Cl Br I	41	—CH_3 =CH_2 ≡CH	52	Si Ge
04	=N—H	15		32	F	42	Md (неметалл)	53	B
05	=N—R	16		33		43	Met (металл)	54	Se Te
06	≡N	21	—SH	34		44	 —C≡X (в ароматических системах)		
07		22	—SR	35		45	—C≡C— (—C≡C—)$_n$		

Не все циклические системы могут быть закодированы по правилам ФКСП, для сложных полициклических структур в качестве кода используется номер из специального списка исключений.

В отличие от линейных циклические дескрипторы описываются буквенно-цифровым кодом переменной длины.

Пример ФКСП-нотации структуры противопротозойного препарата нифуроксима [62]:

5-nitrofuran-2-carbaldehyde oxime



5,06Q1
0100131
0100151
0100351
0164441
0500111
0501351

Первый дескриптор в данном ФКСП-описании циклический, он кодирует пятичленный сопряженный цикл с 6 π-электронами и гетероатомом кислородом в положении 1. Остальные дескрипторы линейные, все они имеют однотипную структуру. Например, последний дескриптор кодирует атом азота =N—R (код 05), соединенный цепочкой из 1 углеродного атома с гетероароматической системой в положении 2 (код 35), причем в цепи есть сопряжение.

ФКСП имеет сложный алфавит и систему словообразования, у него отсутствует синтаксис, нет правил кодирования сложных полициклических структур, он является неоднозначным языком.

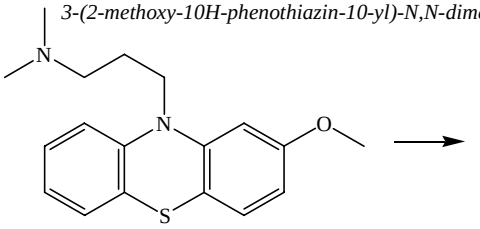
Язык ФКСПм [70] — модификация ФКСП, разработанная в 2000 г., лишена некоторых недостатков своего прототипа. В ФКСПм первичная структура языка практически полностью сохранена, изменения коснулись следующего. Расширен список дескрипторных центров: введены отдельные ДЦ для Br и I (коды 48, 49), новые ДЦ атомов углерода в алифатических циклах в состоянии sp^3-гибридизации (коды 41, 42); как дескрипторные центры нотируются отдельные «висячие» алифатические циклы. При формировании

линейных дескрипторов между ДЦ ищутся все пути, а не только кратчайшие. Полициклические структуры кодируются в виде совокупности цепочек из 2 или 3 базисных циклов (размером не более 8 атомов).

Введены метки дескрипторов в виде номеров дескрипторных центров или циклических дескрипторов, что позволяет объединять первичные дескрипторы в более крупные структурные фрагменты. Но несмотря на все улучшения, ФКСПм остается неоднозначным языком.

Предусмотрена возможность расширения алфавита (за счет введения новых видов дескрипторных центров), что позволяет создавать версии ФКСПм, учитывающие специфику решаемой задачи.

Пример ФКСПм-кодирования структуры нейролептика метоксипромазина [70]:



В приведенном примере первые три дескриптора циклические: первый кодирует шестичленный сопряженный цикл с 6 π-электронами; второй — бициклическую сопряженную систему из двух конденсированных по ребру шестичленных циклов с 12 π-электронами и двумя гетероатомами, азотом и серой, в положениях 6 и 9; третий — трициклическую сопряженную систему, в которой третий шестичленный цикл конденсирован через два ребра (код «B») с конденсированными по ребру двумя шестичленными циклами, система имеет 16 π-электронов и два гетероатома, азот и серу, в положениях 6 и 13.

Остальные дескрипторы линейные, в ФКСПм они имеют такую же структуру, как в ФКСП. Например, последний дескриптор кодирует атом серы —SR (код 22), непосредственно присоединенный к бензольному кольцу (код 33), причем вдоль связи есть сопряжение.

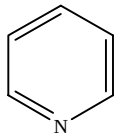
ФКСПм, как и первый вариант ФКСП, используется в различных программах ДСМ-метода (Джон Стюарт Милль, основатель индуктивной логики) [71].

Как интересное развитие ФКСПм отметим попытку создания на его основе языка для описания трехмерных моделей молекул 3DRCPC (3D Reaction Center Pair Code), в котором пары дескрипторных центров дополнительно нотируются параметрами расстояния и угловыми параметрами [72].

В 1981 г. для системы OREX был создан **язык МСДЦ** (матрица связности дескрипторных центров) и разработанный на его основе **язык ФКСП-2** [66, 73], в которых первичный вариант ФКСП претерпел существенные изменения, в основном за счет применения теории графов.

Структура языка кардинально изменилась, в частности, были исключены циклический тип дескриптора и линейные дескрипторы замещения в циклах. Введен новый тип дескрипторных центров — циклические ДЦ, в качестве которых выступают все базисные циклы размером не более 9 атомов и все прочие циклы размером не более 7 атомов. Циклический ДЦ кодируется двузначным числом: первая цифра указывает размер цикла, а вторая — число π-электронов в циклической сопряженной системе (0, если сопряжение связей отсутствует).

Введено новое понятие — индекс вхождения дескрипторных центров как число общих атомов у двух ДЦ, один из которых обязательно обозначает цикл или систему кратных связей. Индекс вхождения кодируется отрицательным числом. Например, в пиридине



индекс вхождения двух дескрипторных центров «шестичленный сопряженный цикл» (код 66) и —N= (код 05) равен −1.

Значительно изменен основной список ДЦ:

— исключен ДЦ оксониевого кислорода (код 16);

— исключены все ДЦ, связанные с описанием циклических структур (коды 33—37, 40, 44);

— введены отдельные ДЦ тройной связи и кумулированных двойных связей (коды 45, 47);

— изменен код атома фосфора (50);

— в ДЦ с кодом 51 добавлен висмут;

— ДЦ неметалла (код 43), бора (код 53), селена и теллура (код 54) объединены в один ДЦ «другие металлоиды» (код 54).

В отличие от ФКСП длина цепи между двумя ДЦ в МСДЦ указывается числом σ-связей, причем рассматриваются все пути, а не только кратчайшие — длиной до 10 связей включительно.

Используемый в ФКСП бинарный индекс сопряжения связей преобразован в индекс связи, который может принимать следующие значения:

0 или 1, если все связи в цепи одинарные;

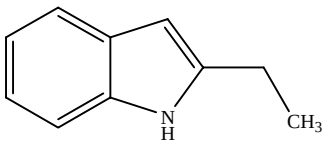
2 или 3, если в цепи есть кратные связи;

4 или 5, если цепь частично проходит по ароматической системе;

7, если цепь полностью проходит по ароматической системе.

Четные цифры в индексе связи соответствуют отсутствию в цепи сопряжения связей, нечетные — их наличию.

Например, в 2-этилиндоле

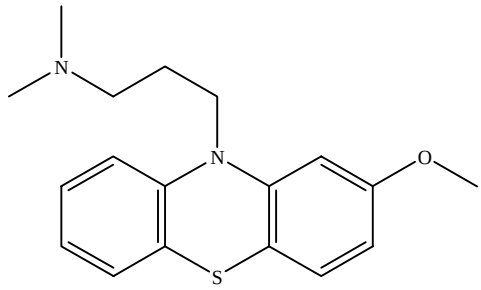


индекс связи между двумя дескрипторными центрами —NH— (код 04) и $—CH_3$ (код 41) равен 5, а между двумя дескрипторными центрами «шестичленный сопряженный цикл» (код 66) и —NH— (код 04) равен 7.

В качестве обобщенной характеристики пути между двумя дескрипторными центрами в МСДЦ используется индекс цепи, который представляет собой трехзначное число, первые две цифры которого отображают длину цепи или индекс вхождения, а третья — индекс связи (для индекса вхождения отсутствует).

В качестве основной формы представления структуры соединения в формате МСДЦ используется матрица связности, в которой диагональными элементами являются ДЦ, а недиагональными — индекс цепи (для индекса вхождения отсутствует).

Приведенная выше структура нейролептика метоксипромазина в формате МСДЦ выглядит следующим образом [73]:



№ ДЦ	1	2	3	4	5	6	7	8	9	10
1	12	045	084	055	094	094	010	055	011	035
2		03	040	035	050	050	054	011	011	−1
3			03	074	010	010	094	050	050	040
4				22	084	084	064	011	011	−1
5					41	020	104	060	060	050
6						41	104	060	060	050
7							41	064	020	080
8								66	021	−2
9									66	−2
10										68

МСДЦ является однозначным языком, поскольку форма представления структуры матричная.

МСДЦ легко разложить на тройки вида (e_{ii}, e_{ij}, e_{jj}), состоящие из двух диагональных и одного недиагонального элементов. В результате получаем линейное описание структурной формулы соединения, которое состоит из совокупности дескрипторов, имеющих общую структуру $ДЦ_1$—(индекс цепи)—$ДЦ_2$. Эта форма первичного представления МСДЦ получила название **ФКСП-2**.

Приведем пример. В записанной выше матрице пересечение двух строк с номерами 1 и 2 кодирует фрагмент структуры, состоящий из дескрипторных центров —OR (код 12) и —NRR (код 03), соединенных цепочкой связей длиной 4, причем цепь частично проходит по ароматической системе и в ней есть сопряжение (индекс цепи 045). Пересечение двух строк с номерами 2 и 10 кодирует шестичленный сопряженный цикл с 8 -электронами (код 68), включающий атом азота —NRR (код 03, индекс вхождения −1).

При формировании полного описания структур дескрипторы ФКСП-2 дополняются подграфами, которые образуются в результате попарного пересечения дескрипторов МСДЦ соединений обучающей выборки.

По сравнению с ФКСП язык ФКСП-2 значительно более унифицирован и пригоден для компьютерной обработки, структура языка более однородная.

Минимальным словом в МСДЦ и ФКСП-2 попрежнему остался крупноблочный линейный дескриптор, который имеет фиксированную длину и кодируется семизначным числом. ФКСП-2, как и прототип ФСКП, является неоднозначным языком.

Прогноз биологической активности в системе OREX (Оптимизированный Распознающий алгоритм конструирования лекарств — Экспертная система, ORacl-EXpert) выполняется на основе статистического метода с использованием значимых подструктурных фрагментов [66].

Язык QL информационной технологии «Микрокосм»

Язык QL (QSAR Language) является одним из главных компонентов информационной технологии прогноза свойств органических соединений «Микрокосм» [2, 74—80].

Первая версия QL появилась в 1988 г. как дальнейшее развитие языков ФКСП и ФКСП-2 и поэтому носила название ФКСП-3 [76—78]. Поскольку новый язык по целому ряду особенностей принципиально отличался от своих предшественников, его более поздние версии получили названия QL-1 — используется с 2000 г. [79] и QL-2 — разработан в 2004 г. [2, 74, 75, 80].

Отметим наиболее существенные отличия языка QL от языков ФКСП и ФКСП-2. Алфавит языка QL представляет собой логически завершенную самостоятельную лингвистическую конструкцию. Язык детализирован и химически содержателен, «буквы» выступают в качестве самостоятельных семантических элементов; ранее минимальным элементом языка являлся дескриптор достаточно большого размера (слово). Язык QL имеет синтаксис, в том числе для порождения крупноблочных дескрипторов; ранее грамматика состояла только из правил словообразования. Структура языка развитая, иерархическая многоуровневая, что позволяет формировать комплексное описание структуры соединений избыточным множеством дескрипторов разной сложности. Унифицирована обработка молекулярных систем с кратными связями. Существенно расширена нотация распределения электронной плотности в дескрипторах. Появилась возможность обработки ансамблей углеродных атомов различного типа.

Язык QL основан на следующих базовых концепциях [2, 74, 81—83].

• *Химическая динамическая система высокой сложности* — совокупность очень большого числа индивидуальных химических соединений, находящихся в пространстве ограниченного объема, которые взаимодействуют между собой и внешней окружающей средой и отделены от среды и друг от друга одной или

несколькими полупроницаемыми поверхностями [81]. В сложной химической системе все внутренние компоненты различного уровня сложности связаны между собой очень большим числом внутренних связей; для нее невозможно построение сколько-нибудь точной многопараметровой аналитической модели. К таким системам, в частности, относятся все биологические системы.

Активность химического соединения есть его способность вызывать изменение значений одного или нескольких внешних параметров сложной химической системы при введении в нее этого соединения. Сложная химическая система (биологическая система) воспринимает конкретное химическое соединение через совокупность очень большого числа его характеристик, по отдельности малозначимых.

• *Обобщенный образ класса соединений с заданным свойством* — совокупность всех возможных структур соединений, проявляющих данное свойство, описанных совокупностью всех возможных параметров, характеризующих эти соединения [82, 83]. Чем больше структур соединений присутствует в выборке и чем больше параметров используется для их описания, тем более точной и адекватной будет модель обобщенного образа.

• *Мультидескрипторное иерархическое многоуровневое описание структуры химических соединений.* В рамках этой концепции для представления структур соединений предусматривается использование различных по смыслу способов описания: групп параметров, с одновременным разделением этих групп на несколько возрастающих по сложности уровней описания химической структуры, с расширяющейся избыточностью такого описания [74].

Из базовых концепций вытекает несколько важных следствий [2]: структура соединений должна быть описана как можно большим числом параметров; описание структуры должно быть многоуровневым; группы параметров описания должны различаться по физическому смыслу; решающие правила для прогноза активности химических соединений должны включать в себя все параметры описания структуры. Исходя из этого можно сформулировать следующие постулаты языка QL [2].

1. Информационное взаимодействие соединения с химическими системами обеспечивается фрагментами структуры с электронодонорными или электроноакцепторными свойствами.

2. σ-Связи соединения образуют каркас для информационных фрагментов структуры.

3. Влияние информационных фрагментов друг на друга определяется длиной пути и степенью делокализации электронов по каркасу структуры.

4. Активность соединения обусловлена комплексным воздействием на химическую систему всех информационных фрагментов, простых и составных.

Итак, *язык QL представляет собой специализированный мультидескрипторный иерархический многоуровневый язык описания структуры химических соединений с фрагментарной подструктурной нотацией [2, 74, 80].*

Алфавит QL постулирован и задается тремя типами элементарных дескрипторов: структурными дескрипторами, дескрипторами длины и дескрипторами связи.

Структурный дескриптор (СД) — фрагмент структуры соединения, обладающий достаточно выраженными электронодонорными или электроноакцепторными свойствами и представляющий собой «неводородный» атом с учетом числа соседей, типов связей и формального заряда или элементарный цикл с учетом сопряжения и ароматичности.

Подалфавит структурных дескрипторов охватывает все элементы периодической системы и состоит из 4352 видов дескрипторов: 378 гетероатомных, 11 углеродных и 3963 циклических. Символ гетероатомного или углеродного СД формируется из символа химического элемента, числа атомов водорода, типов связей и формального заряда; неконцевой sp^3-гибридизованный углерод обозначается >C(<), ароматический —C(Ar)<. Символ циклического СД формируется из символа «Cyc», символа «Ar» (если цикл ароматический), размера цикла от 03 до 99 и числа π-электронов сопряжения (если цикл сопряженный, но не ароматический). При формировании циклических дескрипторов рассматриваются только минимальные базисные циклы.

Помимо символа каждый вид структурного дескриптора имеет соответствующий ему кодовый номер.

Подалфавит структурных дескрипторов представлен в табл. 3.

Дескриптор длины (ДД) — число σ-связей между двумя СД либо число общих атомов у циклического дескриптора и любого другого СД (индекс вхождения, он принимает отрицательные значения). При формировании дескрипторов длины рассматриваются только кратчайшие пути по углеродным цепочкам, при этом гетероатомные СД не пересекаются. Цепочка, которая начинается дескрипторами >C(<) или —C(Ar)<, может оканчиваться только другими СД.

Подалфавит дескрипторов длины состоит из 197 видов дескрипторов.

Дескриптор связи (ДС) характеризует тип электронной системы на дескрипторе длины. Он состоит из трех индексов связей и индекса сопряжения.

В индексах связей типы связей обозначаются следующими строчными буквами (если такая связь на ДД одна), либо прописными буквами (если таких связей на ДД несколько):

p, P — кратные (двойные или тройные);
a, A — связи в ароматических системах;
n, N — нековалентные;
«.» — этого типа связей нет.

Индекс сопряжения может принимать значения 1 (при наличии на ДД сопряжения) или 0 (сопряжения на ДД нет).

Подалфавит дескрипторов связи состоит из 54 видов дескрипторов: 2 одноиндексных, 12 двухиндексных, 24 трехиндексных и 16 четырехиндексных.

Помимо символа каждый вид дескриптора связи имеет соответствующий ему кодовый номер.

На основе элементарных дескрипторов в языке QL порождаются составные дескрипторы более высокого уровня сложности (ранга).

Элементарные дескрипторы — это простые дескрипторы 1-го ранга. Дескрипторы, состоящие из 2—4 элементарных дескрипторов, являются простыми

Таблица 3

Структурные дескрипторы языка QL.

Для каждого структурного дескриптора указан его кодовый номер и символ.
Для дескрипторов «Прочие металлы» Mt — символ химического элемента

№	СД	№	СД	№	СД	№	СД	№	СД	№	СД
	Азот		Фосфор	63	I+n	91	Sn+n	120	—Bi<	152	K+
1	—NH2	32	—PH2	64	—At		Свинец	121	—Bi=	153	—Rb
2	>NH	33	>PH	65	—At′	92	—PbH3	122	—>Bi<	154	Rb+
3	=NH	34	=PH	66	At+n	93	>PbH2	123	—>Bi=	155	—Cs
4	—N<	35	—P<		Бор	94	—PbH<	124	>Bi+<	156	Cs+
5	—N=	36	—P=	67	—BH2	95	>Pb<	125	Bi′n		Прочие металлы
6	#N	37	—PH2=	68	>BH	96	>Pb=	126	Bi+n	157—301	Mt
7	—=N=	38	>PH<	69	—B<	97	Pb′n		Селен	158—302	Mt+n
8	—NH3+	39	>PH=	70	—B=	98	Pb+n	127	—SeH	11159—11303	Mt'n
9	>NH2+	40	—>P<	71	>B′<		Мышьяк	128	>Se		Углерод
10	=NH2+	41	—>P=	11072	>B+	99	—AsH2	129	=Se	303	—C+<
11	—NH+<	42	—=P=		Кремний	100	>AsH	131	>Se<	304	—C′<
12	—NH+=	43	—PH3+	72	—SiH3	101	—As<	132	>Se=	305	—CH3
13	>N+<	44	>PH2+	73	>SiH2	102	—As=	133	=>Se=	306	=CH2
14	>N+=	45	—PH+<	74	—SiH<	103	—>As<	134	—Se+<	307	#CH
15	>N′	46	—PH+=	75	>Si<	104	—>As=	135	—Se′	308	—CH=
	Кислород	47	>P+<	76	>Si=	105	>As+<	136	Se+n	309	—C#
16	—OH	48	>P+=	77	Si′n	106	>As′		Теллур	310	>C=
17	>O	49	>P′		Германий	107	>>As′<	137	—TeH	311	=C=
18	=O	50	>P′<	78	—GeH3		Сурьма	138	>Te	312	>C(<)
20	—O+<	51	>P′=	79	>GeH2	108	—SbH2	139	=Te	315	—C(Ar)<
21	—O+=	52	>>P′<	80	—GeH<	109	>SbH	141	>Te<		Циклы
22	—O′		Галогены	81	>Ge<	110	—Sb<	142	>Te=	500—599	Cycnn
	Сера	53	—F	82	>Ge=	111	—Sb=	143	=>Te=	600—10302	Cycnnkk
23	—SH	54	—F′	83	Ge′n	112	—>Sb<	144	—Te+<	10303—10399	CycArnn
24	>S	55	—Cl	84	Ge+n	113	—>Sb=	145	—Te′		Инертные газы
25	=S	56	—Cl′		Олово	114	>Sb+<	146	Te+n	10403	He
27	>S<	57	Cl+n	85	—SnH3	115	>Sb′		Щелочные металлы	10404	Ne
28	>S=	58	−Br	86	>SnH2	116	Sb′n	147	—Li	10405	Ar
29	=>S=	59	−Br'	87	—SnH<	117	Sb+n	148	Li+	10406	Kr
30	—S+<	60	Br+n	88	>Sn<		Висмут	149	—Na	10407	Xe
31	—S′	61	−I	89	>Sn=	118	—BiH2	150	Na+	10408	Rn
		62	−I'	90	Sn′n	119	>BiH	151	—K		

составными дескрипторами 2—4-го рангов. Всего в QL насчитывается 11 типов простых дескрипторов 1—4-го рангов.

Первичное QL-представление структуры формируется в виде списка дескрипторов 4-го ранга, в которых все структурные дескрипторы пронумерованы (в произвольном порядке). В дальнейшем из этих дескрипторов порождаются дескрипторы всех остальных рангов, поэтому дескрипторы 4-го ранга называются также базовыми дескрипторами. Дескрипторы 5-го и более высоких рангов состоят из двух и более базовых дескрипторов и называются сложными дескрипторами.

В словаре QL все элементарные дескрипторы являются самостоятельными словами.

Простейший синтаксис языка QL состоит всего из двух правил.

1. Каждый последующий по рангу простой дескриптор (2—4-го рангов) образуется в результате конъюнкции одного из дескрипторов предыдущего ранга и одного из элементарных дескрипторов.

2. Каждый последующий по рангу сложный дескриптор (рангом 5 и выше) образуется в результате конъюнкции одного из дескрипторов предыдущего ранга и одного из базовых дескрипторов, причем такое объединение происходит через общий структурный дескриптор, имеющий один и тот же порядковый номер в обоих дескрипторах.

Наличие нумерации структурного дескриптора в базовом дескрипторе и правило порождения сложных

дескрипторов дают возможность полностью воссоздать структурную формулу соединения. Таким образом, QL является однозначным языком.

Строка из чисел простых дескрипторов практически однозначно соответствует структурной формуле соединения, поэтому в версии информационной технологии «Микрокосм-3.2» для прогноза используются именно 11 типов простых дескрипторов языка QL.

Каждый QL-дескриптор имеет две формы представления: внутреннюю (в виде одного или нескольких чисел — для ускорения компьютерной обработки данных и внешнюю в виде привычного «химизированного» символа — для удобного восприятия пользователем. После небольших предварительных пояснений символьная форма QL-дескрипторов совершенно однозначно воспринимается химиком и практически не требует расшифровки.

Обработка структуры соединения происходит в следующей последовательности:

— выделение структурных дескрипторов;

— нахождение циклов;

— определение кратчайших путей и индексов вхождения;

— анализ сопряжения и типа связей;

— выделение дескрипторов ароматических циклов;

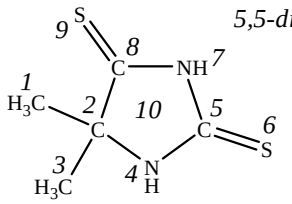
— генерация базовых дескрипторов;

— генерация элементарных дескрипторов;

— генерация простых составных дескрипторов.

При построении решающих правил для всех соединений обучающей выборки рассчитываются матрицы дескрипторов, которые содержат полные QL-описания активных и неактивных структур и представляют собой модели обобщенных образов классов активных и неактивных соединений.

В табл. 4 и 5 приведен пример QL-кодирования структуры противоэпилептического препарата тиомедана



5,5-dimethylimidazolidine-2,4-dithione

В формуле для всех структурных дескрипторов указаны их порядковые номера.

Поясним систему QL-кодирования на примере некоторых дескрипторов из табл. 5:

(—CH3 3) — метильная группа в цепочке длиной 3 с произвольными типами связей и наличием или отсутствием сопряжения;

(=S —CH3) — тионовая группа, соединенная с метильной группой произвольной цепочкой;

(Cyc05 p..0) — пятичленный несопряженный цикл, присоединенный к цепочке произвольной длины, в которой есть одна кратная связь и нет сопряжения;

(2 ...0) — цепочка длиной 2, в которой все связи одинарные и отсутствует сопряжение;

(>NH 3 —CH3) — вторичная аминогруппа, соединенная с метильной группой цепочкой длиной 3 с произвольными типами связей и наличием или отсутствием сопряжения;

Таблица 4

Базовые QL-дескрипторы структуры тиомедана.

N_1, N_2 — номера СД, образующие базовый дескриптор

$СД_1$	ДД	$СД_2$	ДС	N_1	N_2	$СД_1$	ДД	$СД_2$	ДС	N_1	N_2
>NH	1	>C=	...1	7	8	=S	1	>C=	p..1	6	5
>NH	1	>C=	...1	7	5	=S	1	Cyc05	p..0	9	10
>NH	1	>C=	...1	4	5	=S	1	Cyc05	p..0	6	10
>NH	1	>C(<)	...0	4	2	=S	2	>C(<)	p..0	9	2
>NH	2	>NH	...1	4	7	=S	3	—CH3	p..0	9	3
>NH	2	=S	p..1	7	9	=S	3	—CH3	p..0	9	1
>NH	2	=S	p..1	7	6	—CH3	1	>C(<)	...0	3	2
>NH	2	=S	p..1	4	6	—CH3	1	>C(<)	...0	1	2
>NH	2	—CH3	...0	4	3	—CH3	1	Cyc05	...0	3	10
>NH	2	—CH3	...0	4	1	—CH3	1	Cyc05	...0	1	10
>NH	2	>C=	...0	4	8	—CH3	2	—CH3	...0	1	3
>NH	2	>C(<)	...0	7	2	—CH3	2	>C=	...0	3	8
>NH	3	=S	p..0	4	9	—CH3	2	>C=	...0	1	8
>NH	3	—CH3	...0	7	3	>C=	1	>C(<)	...0	8	2
>NH	3	—CH3	...0	7	1	>C=	−1	Cyc05	...0	8	10
>NH	−1	Cyc05	...0	7	10	>C=	—	Cyc05	...0	5	10
>NH	−1	Cyc05	...0	4	10	>C(<)	−1	Cyc05	...0	2	10
=S	1	>C=	p..1	9	8						

Таблица 5

Полное QL-описание структуры тиомедана.

K — число дескрипторов в структуре

СД$_1$	ДД	СД$_2$	ДС	*K*	СД$_1$	ДД	СД$_2$	ДС	*K*	СД$_1$	ДД	СД$_2$	ДС	*K*
>NH				2	—CH3			...0	12	—CH3	2		...0	6
=S				2	—CH3			p..0	2	—CH3	3		...0	2
—CH3				2	>C=			...0	6	—CH3	3		p..0	2
>C=				2	>C=			...1	3	>C=	1		...0	1
>C(<)				1	>C=			p..1	2	>C=	1		...1	3
Cyc05				1	>C(<)			...0	6	>C=	1		p..1	2
	1			13	>C(<)			p..0	1	>C=	2		...0	3
	2			12	Cyc05			...0	7	>C=	−1		...0	2
	3			5	Cyc05			p..0	2	>C(<)	1		...0	4
	−1			5		1		...0	6	>C(<)	2		...0	1
			...0	20		1		...1	3	>C(<)	2		p..0	1
			...1	4		1		p..0	2	>C(<)	−1		...0	1
			p..0	6		1		p..1	2	Cyc05	1		...0	2
			p..1	5		2		...0	7	Cyc05	1		p..0	2
>NH	1			4		2		...1	1	Cyc05	−1		...0	5
>NH	2			9		2		p..0	1	>NH		>NH	...1	1
>NH	3			3		2		p..1	3	>NH		=S	p..0	1
>NH	−1			2		3		...0	2	>NH		=S	p..1	3
=S	1			4		3		p..0	3	>NH		—CH3	...0	4
=S	2			4		−1		...0	5	>NH		>C=	...0	1
=S	3			3	>NH	1	>C=		3	>NH		>C=	...1	3
—CH3	1			4	>NH	1	>C(<)		1	>NH		>C(<)	...0	2
—CH3	2			6	>NH	2	>NH		1	>NH		Cyc05	...0	2
—CH3	3			4	>NH	2	=S		3	=S		—CH3	p..0	2
>C=	1			6	>NH	2	—CH3		2	=S		>C=	p..1	2
>C=	2			3	>NH	2	>C=		1	=S		>C(<)	p..0	1
>C=	−1			2	>NH	2	>C(<)		1	=S		Cyc05	p..0	2
>C(<)	1			4	>NH	3	=S		1	—CH3		—CH3	...0	1
>C(<)	2			2	>NH	3	—CH3		2	—CH3		>C=	...0	2
>C(<)	−1			1	>NH	−1	Cyc05		2	—CH3		>C(<)	...0	2
Cyc05	1			4	=S	1	>C=		2	—CH3		Cyc05	...0	2
Cyc05	−1			5	=S	1	Cyc05		2	>C=		>C(<)	...0	1
>NH		>NH		1	=S	2	>C(<)		1	>C=		Cyc05	...0	2
>NH		=S		4	=S	3	—CH3		2	>C(<)		Cyc05	...0	1
>NH		—CH3		4	—CH3	1	>C(<)		2	>NH	1	>C=	...1	3
>NH		>C=		4	—CH3	1	Cyc05		2	>NH	1	>C(<)	...0	1
>NH		>C(<)		2	—CH3	2	—CH3		1	>NH	2	>NH	...1	1
>NH		Cyc05		2	—CH3	2	>C=		2	>NH	2	=S	p..1	3
=S		—CH3		2	>C=	1	>C(<)		1	>NH	2	—CH3	...0	2
=S		>C=		2	>C=	−1	Cyc05		2	>NH	2	>C=	...0	1
=S		>C(<)		1	>C(<)	−1	Cyc05		1	>NH	2	>C(<)	...0	1
=S		Cyc05		2	>NH	1		...0	1	>NH	3	=S	p..0	1
—CH3		—CH3		1	>NH	1		...1	3	>NH	3	—CH3	...0	2
—CH3		>C=		2	>NH	2		...0	4	>NH	−1	Cyc05	...0	2
—CH3		>C(<)		2	>NH	2		...1	2	=S	1	>C=	p..1	2
—CH3		Cyc05		2	>NH	2		p..1	3	=S	1	Cyc05	p..0	2
>C=		>C(<)		1	>NH	3		...0	2	=S	2	>C(<)	p..0	1
>C=		Cyc05		2	>NH	3		p..0	1	=S	3	—CH3	p..0	2
>C(<)		Cyc05		1	>NH	−1		...0	2	—CH3	1	>C(<)	...0	2
>NH			...0	9	=S	1		p..0	2	—CH3	1	Cyc05	...0	2
>NH			...1	5	=S	1		p..1	2	—CH3	2	—CH3	...0	1
>NH			p..0	1	=S	2		p..0	1	—CH3	2	>C=	...0	2
>NH			p..1	3	=S	2		p..1	3	>C=	1	>C(<)	...0	1
=S			p..0	6	=S	3		p..0	3	>C=	−1	Cyc05	...0	2
=S			p..1	5	—CH3	1		...0	4	>C(<)	−1	Cyc05	...0	1

(>NH 2 ...0) — вторичная аминогруппа в цепочке длиной 2, в которой все связи одинарные и отсутствует сопряжение;

(>NH >NH ...1) — две вторичные аминогруппы, соединенные цепочкой произвольной длины, в которой все связи одинарные и есть сопряжение;

(>C= -1 Cyc05 ...0) — замещенный углерод с двойной связью, входящий в пятичленный несопряженный цикл.

Языку QL присущи следующие полезные свойства, которые позволяют считать его хорошим языком для 2D-QSAR [2].

1. Полнота (универсальность): QL обеспечивает возможность обработки структур неполимерных органических соединений, содержащих атомы любых химических элементов.

2. Непротиворечивость: элементы QL-описания находятся между собой в адекватном смысловом соответствии — цикл, нотированный как не содержащий сопряженной π-электронной системы в одном дескрипторе, не будет определен как ароматический в другом.

3. Содержательность: слова и фразы QL (фрагментарные подструктурные дескрипторы) отражают химические и биологические особенности моделируемой системы.

4. Мультидескрипторность: QL-нотация структуры производится разными по физико-химическому смыслу группами дескрипторов. Так, структурный дескриптор отражает локальные геометрические, электронные и липофильные характеристики, дескриптор длины — интегральные геометрические характеристики, дескриптор связи — интегральные электронные характеристики.

5. Многоуровневость: QL-нотация структуры производится разными по уровню сложности типами дескрипторов.

6. Иерархичность: очередной по сложности уровень QL-дескрипторного описания структуры порождается из предыдущего.

7. Избыточность: структура нотируется большим числом QL-дескрипторов, семантически связанных друг с другом. Например, структура тиомедана описывается 165 видами дескрипторов 1—4-го рангов, поразному отражающих особенности его строения.

8. Интерпретируемость и однозначность: по символу QL-дескриптора легко воспроизвести структурную формулу соответствующего ему фрагмента любой сложности — вплоть до полной структурной формулы соединения.

9. Линейность: все QL-дескрипторы записываются в виде строки.

10. Наглядность: символы QL-дескрипторов легко воспринимаются химиками.

Программный комплекс информационной технологии «Микрокосм-3.2» выполняет прогноз свойств по 11 уровням QL-описания структуры соединений четырьмя различными методами (вероятностным методом Байеса, геометрическим методом расстояния до центров классов, геометрическим методом ближайшего соседа, смешанным методом локального распределения); последующее обобщение спектра из 44 прогнозных оценок производится с использованием трех стратегий голосования: консервативной, нормальной и рисковой [2, 74, 75, 84, 85].

Все версии языка QL в составе программных комплексов «Микрокосм» успешно использовались для прогноза самых разных свойств органических соединений, как биологических, так и небиологических: канцерогенной опасности структурно разнородных соединений [79]; канцерогенной опасности промышленных ускорителей вулканизации резин [86, 87]; анти-ВИЧ активности [88]; 15 видов фармакологической активности многокомпонентного ветеринарного лекарственного препарата растительного происхождения [84, 89, 90]; 3 видов фармакологической активности новых гетероциклических соединений [74]; 20 видов фармакологической активности структурно разнородных соединений [75]; 15 видов свойств (небиологических) модифицирующих добавок для пяти различных полимеров [85]; 4 вида свойств технологических добавок в резиновые смеси [91].

Заключение

Большинство современных QSAR-систем для описания химической структуры использует непрерывные числовые параметры, например, значения физико-химических свойств или параметры, полученные в результате 3D-моделирования. Довольно часто эти параметры рассчитываются после предварительного структурного фрагментирования, непосредственные результаты которого в итоговые прогнозные зависимости не включаются.

Такие работы (как и многие работы, выполненные только на основе фрагментарных нотаций), как правило, не содержат детального описания собственно методики подструктурного кодирования. Вероятно, предполагается, что этот классический способ и так всем хорошо известен. В результате в большинстве публикаций описывается, *что* получено, но не описывается, *как* получено. Настоящая статья представляет собой попытку показать, как именно обрабатываются химические структуры с помощью различных QSAR-ориентированных языков фрагментарного подструктурного кодирования.

Возможности этих языков еще далеко не исчерпаны. Их прогностическая мощность может быть значительно увеличена: путем дополнения фрагментных подструктурных дескрипторов «нотационными» дескрипторами, описывающими локальные распределения тех или иных физико-химических свойств [2]; как результат детализации алфавита с включением новых данных о ранее экзотических соединениях (например, соединения двухкоординированного германия — гермилены [92]) и т.д.

В завершение статьи авторы выражают надежду, что представленная здесь информация о принципах и

методах фрагментарного кодирования структуры органических соединений будет способствовать активизации работ по использованию этих методов для компьютерного прогноза биологической активности.

ЛИТЕРАТУРА

1. *Бутлеров А.М.* Сочинения. В 3 т. М.: Изд-во АН СССР, 1953—1958. Т. 1, с. 70, 73.
2. *Васильев П.М.* Молекулярное моделирование в химии, биологии и медицине. Тез. докл. II Росс. школы-конференции (Саратов, 13—16 окт. 2004 г.) Саратов: СГУ, 2004, с. 8—14.
3. *Голендер В.Е., Розенблит А.Б.* Вычислительные методы конструирования лекарств. Рига: Зинатне, 1978, 238 с.
4. *Барлоу Р.* Введение в химическую фармакологию. М.: Издатинлит, 1959, с. 54.
5. *Wiswesser W.J.* A Line-Formula Chemical Notation. N. Y.: Thomas Y., Crowell Co., 1954.
6. *Smith E.G.* The Wiswesser Line-Formula Chemical Notation. N. Y.: McGrow-Hill, 1968.
7. *Стьюпер Э., Брюгер У., Джурс П.* Машинный анализ связи химической структуры и биологической активности. М.: Мир, 1982, с. 74.
8. *Dyson G.M., Lynch M.F., Morgan H.L.* Inform. Storage and Retrieval., 1968, № 4, p. 27—83.
9. *Горелик А.Л., Скрипкин В.А.* Методы распознавания. Учеб. пособие. 2-е изд. М.: Высшая школа, 1984, с. 154.
10. *Toler L.* Division of Chemical Information: 222-nd ACS National Meeting (Chicago, IL, USA, 26—30 Aug. 2001). Chicago, 2001. Chemical Identifiers: Names and Structures: Kurt Loening Memorial Symposium. Present. 6.
11. *Cramer R.D., Redl G., Berkoff C.E.* J. Med. Chem., 1974, v. 17, № 5, p. 533—535.
12. *Adamson G.W., Bawden D.* J. Chem. Inf. Comput. Sci., 1976, v. 16, № 3, p. 161—169.
13. *Gluck D.J.* J. Chem. Doc., 1965, v. 5, № 1, p. 43—51.
14. *Morgan H.L.* Ibid., 1965, v. 5, № 2, p. 107—113.
15. CTfile Formats: December 1998. San Leandro, CA, USA: MDL Information Systems, Inc., 1998, 96 p.
16. *Weininger D.* J. Chem. Inf. Comput. Sci., 1988, v. 28, № 1, p. 31—36.
17. *Weininger D., Weininger A., Weininger J.L.* Ibid., 1989, v. 29, № 2, p. 97—101.
18. *Weininger D.* Ibid., 1990, v. 30, № 3, p. 237—243.
19. Chemoinformatics: A Textbook. Eds. J. Gasteiger, T. Engel. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2003, 649 p.
20. *Convard T., Dubost J.P., Lesollev H., Kummer E.* Quant. Struct. Act. Rel., 1994, v. 13, № 1, p. 34—37.
21. *Bone R.G.A., Firth M.A., Sykes R.A.* J. Chem. Inf. Comput. Sci., 1999, v. 39, № 5, p. 846—860.
22. *Xue L., Godden J.W., Bajorat J.* Ibid., 1999, v. 39, № 5, p. 881—886.
23. *Barnard J.M., Jochum C.J., Welford S.M.* Chemical Structure Information System: Interfaces Communication and Standards. Ed. W.A. Warr. ACS Symposium Series № 400. Washington, DC: ACS, 1989, p. 76—81.
24. *Ash S., Cline M.A., Homer R.W., Hurst T., Smith G.B.* J. Chem. Inf. Comput. Sci., 1997, v. 37, № 1, p. 71—79.
25. *Bawden D., Devon T.K., Jackson F.T., Wood S., Lynch M.F., Willett P.* Ibid., 1979, v. 19, № 2, p. 90—93.
26. Derwent World Patents Index: CPI Chemical Indexing User Guide. Prep.: Bryan P. Revis. Ed. 1. London: Derwent Information, 2000, 278 p.
27. *Раевский О.А.* Успехи химии, 1999, т. 68, № 6, с. 555—576.
28. *Todeschini R., Consonni V.* Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry. V. 11. Eds. R. Mannhold, H. Kubinyi, H. Timmerman. Weinheim: Wiley-VCH, 2000.
29. *Klopman G.* J. Am. Chem. Soc., 1984, v. 106, № 24, p. 7315—7321.
30. *Klopman G.* Quant. Struct. Act. Rel.,1991, v. 10, № 2, p. 176—184.
31. *Graham C., Gealy R., Macina O.T., Karol M.H., Rosenkranz H.S.* Ibid., 1996, v. 15, № 3, p. 224—229.
32. *Meylan W.M., Howard P.H.* J. Pharm. Sci., 1995, v. 84, № 1, p. 83—92.
33. Применение математических методов для анализа связи структура — пестицидная активность. Ч. 3. Сост.: Валуева Л.Н., Зацепин В.М. Обзорн. инф. сер. «Химические средства защиты растений». М.: НИИТЭХИМ, 1987, 51 с.
34. *Лазуткин Е.Ю., Мотовилов М.Б., Семенов В.Д., Нигматуллин Р.С., Осипов А.Л., Зацепин В.М.* ВАТОХ. Первая всесоюзная конференция по теоретической органической химии. Тез. докл. (Волгоград, 1991 г.). Волгоград, 1991, Е-44, с. 539.
35. *Зацепин В.М., Осипов А.Л., Семенов Р.Д.* Автометрия, 1995, № 5.
36. *Willett P.* Cur. Opin. Biotech., 2000, v. 11, № 1, p. 85—88.
37. *Кадыров Ч.Ш., Тюрина Л.А., Симонов В.Д., Семенов В.А.* Машинный поиск химических препаратов с заданными свойствами. Ташкент: Фан, 1989, 164 с.
38. *Тюрина Л.А., Каримова Ф.С., Кирлан А.В., Кирлан С.А., Лукманова А.Л., Шагалеева З.Р., Гильмханова В.Т., Валитов Р.Б., Давыдов А.М.* Агрохимия, 2002, № 3, с. 35—41.
39. *Соломинова Т.С., Пилюгин В.С., Тюрин А.А., Кирлан А.В., Тюрина Л.А.* Хим.-фарм. ж., 2004, т. 38, № 8, с. 20—24.
40. *Moriguchi I., Hirono S., Matsushita Y., Nakagome I.* Chem. Pharm. Bull., 1992, v. 40, p. 930—934.
41. *Moriguchi I., Hirono S., Liu Q., Nakagome I.* Quant. Struct. Act. Rel., 1992. v. 11, № 4, p. 325—331.
42. *King R.D., Muggleton S.H., Srinivasan A., Sternberg M.J.E.* Proc. Natl. Acad. Sci. USA, 1996, v. 93, № 1, p. 438—442.
43. *King R.D., Srinivasan A.* Environ. Health Perspect., 1996, v. 104, Suppl. 5, p. 1031—1040.
44, *Varmuza K., Scsibrany H.* J. Chem. Inf. Comput. Sci., 2000, v. 40, № 2, p. 308—313.
45. *Varmuza K., Demuth W., Karlovits M., Scsibrany H.* Croat. Chem. Acta, 2005, v. 78, № 2, p. 141—149.
46. *Scsibrany H., Karlovits M., Demuth W., Müller F., Varmuza K.* Chemom. Intell. Lab. Syst., 2003, v. 67, p. 95—108.
47. *Баскин И.И., Палюлин В.А., Зефиров Н.С.* ВАТОХ. Первая всесоюз. конф. по теоретической органической химии. Тез. докл. (Волгоград, 29 сен.—05 окт. 1991 г.). Волгоград, 1991, Е-62, с. 557.
48. *Palyulin V.A., Baskin I.I., Petelin D.E., Zefirov N.S.* QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Application. Eds. F. Sanz, J. Giraldo, F. Manaut. Barcelona: Prous Science Publishers, 1995, p. 51—52.
49. *Zefirov N.S., Palyulin V.A.* J. Chem. Inf. Comput. Sci., 2002, v. 42, № 5, p. 1112—1122.
50. *Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С.* Ж. прикл. химии, 2003, т. 76, № 12, с. 1966—1970.
51. *Жохова Н.И., Баскин И.И., Палюлин В.А., Зефиров А.Н., Зефиров Н.С.* Ж. структ. химии, 2004, т. 45, № 4, с. 626—635.
52. *Wife R.L., de Bie M.J.F.* CHI Meeting on Combinatorial Chemistry & High Throughput Screening (Osaka, Japan, 7—9 Nov. 1996). Osaka, 1996.

53. *Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J.* Adv. Drug Deliv. Rev., 1997, v. 23, № 1, p. 3—25.
54. *Глориозова Т.А., Филимонов Д.А., Лагунин А.А., Поройков В.В.* Хим.-фарм. ж., 1998, т. 32, № 12, с. 33—39.
55. *Filimonov D., Poroikov V., Borodina Yu., Gloriozova T.* J. Chem. Inf. Comput. Sci., 1999, v. 39, № 4, p. 666—670.
56. *Зефирова О.Н., Зефиров Н.С.* Вестн. Моск. ун-та. Сер. 2. Химия, 2002, т. 43, № 4, с. 251—256.
57. *Filimonov D.A., Poroikov V.V.* The 14th European Symposium on Quantitative Structure-Activity Relationships (Euro QSAR 2002). Abstracts (Bournemouth, UK, September 2002). Bournemouth, 2002.
58. *Naumann T., Lowis D.* The First International Electronic Conference on Synthetic Organic Chemistry (September 1—30, 1997): CD-ROM Edition. Eds. S.-K. Lin, P.-V. Esteban. MDPI, 1999, F0003. http://www.mdpi.org/ecsoc-1.htm.
59. US patent 5751605. Molecular hologram QSAR.
60. US patent 6208942. Molecular hologram QSAR. 27.04.2001.
61. *Cho S.J.* Bull. Korean Chem. Soc., 2005, v. 26, № 1, p. 85—90.
62. Инструкция по кодированию химической структуры фрагментарным кодом суперпозиции подструктур (ФКСП). Минмедпром СССР. НИИ по БИХС. Купавна, 1972, 30 с.
63. *Авидон В.В.* Хим.-фарм. ж., 1974, т. 8, № 8, с. 22—25.
64. *Avidon V.V., Pomerantsev I.A., Golender V.E., Rozenblit A.B.* J. Chem. Inf. Comput. Sci., 1982, v. 22, № 4, p. 207—214.
65. *Авидон В.В., Аролович В.С., Козлова С.П., Пирузян Л.А.* Хим.-фарм. ж., 1978, т. 12, № 6, с. 99—106.
66. *Розенблит А.Б., Голендер В.Е.* Логико-комбинаторные методы в конструировании лекарств. Рига: Зинатне, 1983, 352 с.
67. *Финн В.К.* Итоги науки и техники. Сер. Информатика. М.: ВИНИТИ, 1991, т. 15, с. 54—101.
68. *Погребняк А.В.* Молекулярное моделирование и дизайн биологически активных веществ. Ростов-на-Дону: Издательство СКНЦ ВШ, 2003, 232 с.
69. *Комиссаров И.В.* Элементы теории рецепторов в молекулярной фармакологии. М.: Медицина, 1969.
70. *Блинова В.Г., Добрынин Д.А.* Науч.-техн. информ. Сер. 2, 2000, № 6, с. 14—21.
71. *Блинова В.Г., Добрынин Д.А., Жолдакова З.И., Харчевникова Н.В.* Там же, Сер. 2, 2001, № 10, с. 13—19.
72. *Добрынин Д.А.* Там же, Сер. 2, 2000, № 6, с. 14—21.
73. *Гитлина Л.С., Голендер В.Е., Дрбоглав В.В., Розенблит А.Б., Эйхенберга Р.А., Авидон В.В.* Методы представления и обработки структурной информации для анализа связи структура—активность. Рига, 1981, 75 с. (Препринт АН ЛатвССР, Ин-т орг. синтеза: 101).
74. *Васильев П.М., Спасов А.А., Косолапов В.А., Степанов А.В., Дудченко Г.П.* Информационные технологии в образовании, технике и медицине: Матер. междунар. конф. (Волгоград, 18—22 окт. 2004 г.). ВолгГТУ. Волгоград, 2004, т. 3, с. 180—186.
75. *Васильев П.М., Спасов А.А.* Вестник ВолГМУ, 2005, т. 13, № 1, с. 23—30.
76. *Васильев П.М., Бреслаухов А.Г.* Синтез и применение пестицидов и кормовых добавок в сельскохозяйственном производстве. Тез. докл. регион. науч.-техн. конф. Волгоград, 1988, с. 145.
77. *Vassiliev P.M., Breslaukhov A.G.* Second World Congress of Theoretical Organic Chemists: Abstracts (University of Toronto, Canada, 8—14 July 1990), Toronto, 1990, AA-38.
78. *Бреслаухов А.Г., Васильев П.М.* ВАТОХ. Первая всесоюзная конференция по теоретической органической химии. Тез. докл. (Волгоград, 29 сен.—05 окт. 1991 г.). Волгоград, 1991, А-2, с. 78.
79. *Васильев П.М., Орлов В.В., Дербишер В.Е.* Хим.-фарм. ж., 2000, т. 34, № 7, с. 19—22.
80. *Васильев П.М.* Тез. докл. XI Росс. национ. конгресса «Человек и лекарство» (Москва, 19—23 апр. 2004 г.). М., 2004.
81. *Васильев П.М.* 1-я Росс. электронная конф. по биоинформатике (RECOB-2000), http://www.ibmh.msk.su/recob/ (Москва, 15 мар.—21 апр. 2000). НИИ биомедицинской химии РАМН. G05, 4 с.
82. *Васильев П.М.* ВАТОХ. Первая всесоюзн. конф. по теоретической органической химии. Тез. докл. (Волгоград, 29 сен.—05 окт. 1991 г.). Волгоград, 1991, А-1 (Е-2), с. 77 (с. 497).
83. *Васильев П.М.* Молекулярное моделирование в химии, биологии и медицине. Тез. докл. I Росс. школы-конф. (Саратов, 18—20 сент. 2002 г.). СГУ. Саратов, 2002, с. 19—20.
84. *Васильев П.М., Горлов И.Ф., Юрина О.С.* Молекулярное моделирование. Тез. докл. 3-й Всеросс. конф. (Москва, 15—17 апр. 2003 г.). М., 2003, с. 54.
85. *Но Б.И., Васильев П.М., Зотов Ю.Л., Новаков И.А., Орлов В.В., Дербишер В.Е., Хмелидзе И.А.* Пластические массы, 2003, № 4, с. 27—32.
86. *Старовойтов М.К., Васильев П.М., Рудакова Т.В., Белоусов Е.К., Крякунов М.В., Качегин А.Ф.* Каучук и резина, 2002, № 1, с. 28—31.
87. *Vassiliev P.M., Rudakova T.V., Beloussov E.K., Tsapkova E.V., Ivankina O.M., Tishin O.A.* 16-th International Congress of Chemical and Process Engineering (CHISA 2004): Summaries (Praha, Czech Republic, 22—26 Aug. 2004). Praha, 2004, v. 5 «Systems and Technology». P7.91, p. 1944—1945.
88. *Васильев П.М.* Тез. докл. IX Росс. национ. конгресса «Человек и лекарство» (Москва, 8—12 апр. 2002 г.). М., 2002, с. 592.
89. *Васильев П.М., Горлов И.Ф., Юрина О.С.* Докл. РАСХН, 2002, № 2, с. 55—58.
90. *Горлов И.Ф., Юрина О.С., Васильев П.М.* Там же, 2002, № 5, с. 45—47.
91. *Васильев П.М., Каблов В.Ф., Хортик К.В., Новопольцева О.М.* Каучук и резина, 2001, № 3, с. 22—25.
92. *Землянский Н.Н., Борисова И.В., Кузнецова М.Г., Хрусталев В.Н., Антипин М.Ю., Устынюк Ю.А., Лунин В.В., Eaborn C., Hill M.S., Smith J.D.* Ж. орг. химии, 2003, т. 39, № 4, с. 527—536.