



Уральский
федеральный
университет

имени первого Президента
России Б.Н.Ельцина

Химико-
технологический
институт

Ю. И. НЕЙН
М. Н. ИВАНЦОВА

КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКОЙ ИНФОРМАЦИИ

Учебное пособие

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б. Н. ЕЛЬЦИНА

Ю. И. Нейн, М. Н. Иванцова

КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКОЙ ИНФОРМАЦИИ

Учебное пособие

Рекомендовано методическим советом
Уральского федерального университета в качестве учебного пособия
для студентов вуза, обучающихся по направлениям подготовки
18.03.01 «Химическая технология»,
19.03.01 «Биотехнология»

Екатеринбург
Издательство Уральского университета
2020

УДК 54:004(075.8)
ББК 4+30.2-5-05я73
Н46

Под общей редакцией М. Ф. Костериной

Р е ц е н з е н т ы:

кафедра химии и процессов горения
Уральского института Государственной противопожарной службы
МЧС России

(и. о. начальника кафедры кандидат химических наук
капитан внутренней службы А. В. Кокшаров);
Е. В. Щегольков, кандидат химических наук,
старший научный сотрудник
лаборатории фторорганических соединений
Института органического синтеза УрО РАН

Нейн, Ю. И.

Н46 Компьютерное представление химической информации : учебное пособие / Ю. И. Нейн, М. Н. Иванцова ; под общ. ред. М. Ф. Костериной ; Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет. — Екатеринбург : Изд-во Урал. ун-та, 2020. — 142 с. : ил. — Библиогр.: с. 141. — 30 экз. — ISBN 978-5-7996-3018-8. — Текст : непосредственный.

ISBN 978-5-7996-3018-8

В учебном пособии рассмотрены основные типы представления органических молекул в компьютерном виде: линейные нотации WLN, ROSDAL, SMILES, SLN, InChI и табличные представления — Z-матрицы, MOL-, SDF- и RDF-форматы. Системное изложение материала позволит студентам самостоятельно подготовиться к занятиям и сдаче зачетов и экзаменов.

Для студентов бакалавриата, осваивающих дисциплины «Основы квантовой химии и хемоинформатики», «Компьютерные информационные системы в биотехнологии» и «Компьютерное моделирование состава продуктов питания», а также для студентов магистратуры, аспирантов и научных работников.

УДК 54:004(075.8)
ББК 4+30.2-5-05я73

ОГЛАВЛЕНИЕ

Предисловие	5
Введение	6
1. ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРНОМ ВИДЕ	9
1.1. Единицы измерения информации	9
1.2. Информационно-поисковые языки	13
1.3. Представление органических молекул в компьютерном виде	18
2. ЛИНЕЙНЫЕ НОТАЦИИ.....	19
2.1. Линейная нотация Висвессера (WLN).....	19
2.2. Представление органических структур в формате ROSDAL.....	30
2.3. Спецификация упрощенного представления молекул в строке ввода (SMILES).....	31
2.4. Линейная нотация сибил (SLN)	43
3. ТАБЛИЧНЫЕ ПРЕДСТАВЛЕНИЯ	45
3.1. Представление 3D-структур	45
3.2. Продолжающиеся таблицы (таблицы соединений)	48
3.3. Мол-файл (Mol-file)	54
3.4. Pdb-файлы (база данных по белкам).....	58
3.5. Z-матрица	60
3.6. Теория графов (диаграммы графов)	66
3.6.1. Матрица смежности	68
3.6.2. Матрица расстояний.....	70
3.6.3. Матрица понижения	71
3.6.4. Матрица связей.....	72
3.6.5. Матрица «связь — электрон» (BE-матрица).....	75
3.7. Представление матрицы химических реакций.....	77
3.7.1. SD-файлы	77
3.7.2. RXN-файлы (реакционные файлы).....	77
3.7.3. RD-файлы (файлы «реакция — данные»).....	80
3.7.4. CML (химический язык разметки).....	81
4. НУМЕРАЦИЯ АТОМОВ	83
4.1. Алгоритм Моргана	83

4.2. CANGEN-алгоритм	86
5. МЕЖДУНАРОДНЫЙ ХИМИЧЕСКИЙ ИДЕНТИФИКАТОР (InChI)	90
5.1. Правила InChI	93
5.1.1. Нумерация атомов (Color List)	94
5.1.2. Записи основных слоев	98
5.2. Inchikey — ключ для поиска структуры	104
ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ	108
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	141

ПРЕДИСЛОВИЕ

Возникновение и развитие информационных справочных изданий по химическим структурам резко увеличило поток новой информации, поиск, просмотр и анализ которой в настоящее время немислим без применения автоматизированных информационно-поисковых технологий. В вопросах разработки таких технологий одной из важнейших задач является составление алгоритма ввода структурных формул органических соединений, позволяющих однозначно представлять двухмерную структурную формулу соединения набором строк символов, удобных для ввода в память ЭВМ.

Предлагаемая книга не претендует на исчерпывающую полноту описания всех возможных вариантов представлений молекул в компьютерном виде. Цель пособия — познакомить студентов с основными наиболее распространенными кодировками и научить самостоятельно кодировать и расшифровывать структуры химических соединений.

В учебном пособии рассмотрены основные типы представления органических молекул в компьютерном виде: линейные нотации WLN, ROSDAL, SMILES, SLN, InChI и табличные представления — Z-матрицы, MOL-, SDF- и RDF-форматы. Книга содержит теоретический и практический материал. В заключительном разделе пособия приводятся варианты контрольных заданий, которые могут быть использованы преподавателем для проведения как текущей, так и итоговой аттестации студентов. Дается список литературы для более углубленного изучения линейных и табличных нотаций.

Данное учебное пособие представляет собой коллективный труд преподавателей кафедры технологии органического синтеза химико-технологического института Уральского федерального университета. Авторы выражают глубокую признательность Ю. Ю. Моржерину за ряд ценных советов, данных им при подготовке пособия.

ВВЕДЕНИЕ

Хемоинформатика — это научная дисциплина, возникшая в последние 50 лет в пограничной области между химией и вычислительной математикой. Во второй половине XX в. стало ясно, что во многих областях химии огромный объем информации, накопленный в ходе химических исследований, может быть обработан и проанализирован только с помощью компьютеров. Более того, многие из проблем в химии настолько сложны, что для их решения требуются новые подходы, основанные на применении методов информатики. Исходя из этого были разработаны методы для построения баз данных по химическим соединениям и реакциям для прогнозирования физических, химических и биологических свойств соединений и материалов, для поиска новых лекарственных препаратов, анализа спектральной информации, для предсказания хода химических реакций и планирования органического синтеза.

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 до 90 % сведений о свойствах изучаемого вещества из структурной формулы.

Традиционно используемая и по сей день классическая фрагментация структурных формул соединений по функциональным группам, кратным связям, циклам, ароматическим или конденсированным системам лежала в основе большинства ранних работ

по исследованию соотношений «структура — биологическая активность». Она позволяет выявлять умозрительные эмпирические закономерности, например, такого 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. Формирование во второй половине XX в. 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 небιологических свойств веществ.

В хемоинформатике для внутреннего представления структур химических соединений обычно используются молекулярные графы, которые могут быть при необходимости дополнены информацией о трехмерных координатах атомов, а также о динамике их изменения во времени. Долговременное хранение химической информации и обмен ею между приложениями осуществляется при помощи файлов, организованных в соответствии с типами внешнего представления химической информации.

Простейшим типом внешнего представления структур химических соединений являются линейные нотации в виде строки символов. Исторически первым видом линейных нотаций явилась линейная нотация Висвессера (WLN). В настоящее время наиболее распространенным видом линейных нотаций являются строки SMILES. Кроме того, применяются линейные нотации SLN (Sybyl Line Notation, Tripos, Inc.; содержит также возможность спецификации структур Маркуша), SMARTS (расширение SMILES для поисковых запросов к химическим базам данных), ROSDAL. Для унификации кодировки химических структур в 2005 г. ИЮПАК (Международный союз теоретической и прикладной химии, IUPAC) принял универсальную линейную нотацию InChI и InChIKey.

Второй тип внешнего представления структур химических соединений и реакций между ними основан на непосредственном кодировании матрицы смежности молекулярного графа. Такие распространенные форматы, как MOL, SDF и RDF, которые в настоящее

время являются общепринятыми стандартными для обмена химической информацией, можно считать способами представления в виде текстового файла матрицы смежности молекулярного графа. Этой же цели служат и специфические форматы MOL2, HIN, PCM и др., предназначенные для работы с распространенными программами по молекулярному моделированию.

Наконец, третий тип внешнего представления структур химических соединений основан на технологии XML. Наиболее распространенным языком описания химической информации, опирающимся на эти принципы, является CML.

1. ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРНОМ ВИДЕ

1.1. ЕДИНИЦЫ ИЗМЕРЕНИЯ ИНФОРМАЦИИ

Обычно информация представляет собой последовательность символов. Каждый символ имеет каноническое изображение, которое позволяет однозначно идентифицировать данный символ. Варианты начертания символов задают разные шрифты.

В вычислительных машинах для представления информации используются цепочки байтов. Поэтому для перевода информации из машинного представления в понятный для человека вид необходимы таблицы кодировки символов — таблицы соответствия между символами определенного языка и кодами символов. Их еще называют кодовыми страницами или применяют английский термин *character set* (который иногда сокращают до *charset*).

В ЭВМ применяется двоичная система, т. е. все числа в компьютере представляются с помощью нулей и единиц, поэтому компьютер может обрабатывать только информацию, представленную в цифровой форме.

Для преобразования числовой, текстовой, графической, звуковой информации в цифровую необходимо применить кодирование. Кодирование — это преобразование данных одного типа через данные другого типа. В ЭВМ применяется система двоичного кодирования, основанная на представлении данных последовательностью двух знаков: 1 и 0, которые называются двоичными цифрами (*binary digit* — сокращенно *bit*).

Таким образом, единицей информации в компьютере является один бит, т. е. двоичный разряд, который может принимать значение 0 или 1. Восемь последовательных битов составляют байт. В одном байте можно закодировать значение одного символа из 256 возможных ($256 = 2$ в степени 8). Более крупной единицей информации является килобайт (Кбайт), равный 1024 байтам ($1024 = 2$ в степени 10). Еще более крупные единицы измерения данных: мегабайт,

гигабайт, терабайт (1 Мбайт = 1024 Кбайт; 1 Гбайт = 1024 Мбайт; 1 Тбайт = 1024 Гбайт).

Целые числа кодируются двоичным кодом довольно просто (путем деления числа на два). Для кодирования нечисловой информации используется следующий алгоритм: все возможные значения кодируемой информации нумеруются и эти номера кодируются с помощью двоичного кода. Например, для представления текстовой информации используется таблица нумерации символов или таблица кодировки символов, в которой каждому символу соответствует целое число (порядковый номер). Восемь двоичных разрядов могут закодировать 256 различных символов.

Самой известной таблицей кодировки является код ASCII (Американский стандартный код для обмена информацией). Первоначально он был разработан для передачи текстов по телеграфу, причем в то время он был 7-битовым, т. е. для кодирования символов английского языка, служебных и управляющих символов использовались только 128 7-битовых комбинаций (табл. 1).

Таблица 1

Первые 128 значений кодировочной таблицы ASCII

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
0.	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	TAB	LF	VT	FF	CR	SO	SI
1.	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2.		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3.	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4.	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5.	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6.	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7.	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

При этом первые 32 комбинации (кода) служили для кодирования управляющих сигналов (начало текста, конец строки, перевод каретки, звонок, конец текста и т. д.). При разработке первых

компьютеров фирмы IBM этот код был использован для представления символов в компьютере. Поскольку в исходном коде ASCII было всего 128 символов, для их кодирования хватило значений байта, у которых 8-й бит равен 0. Значения байта с 8-м битом, равным 1, стали использовать для представления символов псевдографики, математических знаков и некоторых символов из языков, отличных от английского (греческого, немецких умляутов, французских диакритических знаков и т. п.).

Существующий стандарт ASCII (8-разрядная система кодирования) содержит две таблицы кодирования — базовую и расширенную. Первая таблица содержит 128 основных символов, в ней размещены коды символов английского алфавита, а во второй таблице кодирования содержатся 128 расширенных символов. Так как в этот стандарт не входят символы национальных алфавитов других стран, то в каждой стране 128 кодов расширенных символов заменяются символами национального алфавита. В настоящее время существует множество таблиц кодировки символов, в которых 128 кодов расширенных символов заменены символами национального алфавита.

Когда стали приспособлять компьютеры для других стран и языков, места для новых символов уже не стало хватать. Для того чтобы полноценно поддерживать помимо английского и другие языки, фирма IBM ввела в употребление несколько кодовых таблиц, ориентированных на конкретные страны. Так, для скандинавских стран была предложена таблица 865 (Nordic), для арабских стран — таблица 864 (Arabic), для Израиля — таблица 862 (Israel) и т. д. В этих таблицах часть кодов из второй половины кодовой таблицы использовалась для представления символов национальных алфавитов (за счет исключения некоторых символов псевдографики).

С русским языком ситуация развивалась особым образом. Очевидно, что замену символов во второй половине кодовой таблицы можно произвести разными способами. Вот и появились для русского языка несколько разных таблиц кодировки символов кириллицы: KOI8-R, IBM-866, CP-1251, ISO-8551-5. Все они одинаково изображают символы первой половины таблицы (от 0 до 127) и различаются представлением символов русского алфавита и псевдографики.

Так, например, кодировка символов русского языка Windows–1251 используется для компьютеров, которые работают под управлением операционной системы Windows (табл. 2). Другая кодировка для русского языка — это KOI8, которая также широко используется в компьютерных сетях и российском секторе интернета (табл. 3).

Таблица 2

Кодировка символов русского языка Windows–1251

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	Ъ	Ѓ	,	ѓ	„	…	†	‡	€	%	Љ	‹	Њ	Ќ	Ў	Ц
9.	ђ	‘	’	“	”	•	–	—		™	љ	›	њ	ќ	ћ	ц
A.		Ў	ў	Ј	Ѡ	Ѓ	Ѕ	§	Ё	©	€	«	¬		®	ї
B.	°	±	І	і	г	µ	¶	·	ё	№	€	»	ј	ѕ	ѕ	ї
C.	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
D.	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
E.	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
F.	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

Таблица 3

Кодировка символов русского языка KOI8-R

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	—		Г	Г	Л	Л	Т	Т	Т	Т	■	■	■	■	■	■
9.	⋮	⋮	⋮		■	•	√	≈	≤	≥		Ј	°	²	•	÷
A.	=		ƒ	ё	г	г	г	г	г	г	ц	ц	ц	ц	ц	ц
B.				Ё			Т	Т	Т	Т	ц	ц	ц	ц	ц	©
C.	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D.	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ
E.	Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F.	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ

Для таких языков, как китайский или японский, 256 символов недостаточно. Кроме того, всегда существует проблема вывода или сохранения в одном файле одновременно текстов на разных языках (например, при цитировании). Поэтому была разработана универсальная кодовая таблица UNICODE, содержащая символы, применяемые в языках всех народов мира, а также различные служебные и вспомогательные символы (знаки препинания, математические и технические символы, стрелки, диакритические знаки и т. д.).

Очевидно, что одного байта недостаточно для кодирования такого большого множества символов. Поэтому в UNICODE используются 16-битовые (2-байтовые) коды, что позволяет представить 65 536 символов. Для совместимости с предыдущими кодировками первые 256 кодов совпадают со стандартом ASCII. UNICODE 6.0, стандарт 2010, может кодировать до 2^{31} (2 147 483 648) позиций. В UNICODE зарезервировано 1 114 112 ($= 2^{20} + 2^{16}$) позиций символов, из которых сейчас используется свыше 100 000 кодовых позиций (109 242 графические и 273 — прочие символы).

1.2. ИНФОРМАЦИОННО-ПОИСКОВЫЕ ЯЗЫКИ

Возникновение и развитие информационных справочных изданий по химическим структурам — от первого реферативного журнала «Pharmazeutisches centralblatt» (1830) и последовавших за ним фундаментальных справочников по органической химии Бейльштейна, Тейльгеймера, Губена до мощных современных реферативных изданий типа «Реферативный журнал “Химия”», «Chemical Abstracts», «Chemical Titles», — резко увеличило поток новой информации, поиск, просмотр и анализ которой в настоящее время немыслим без применения автоматизированных информационно-поисковых технологий. В вопросах разработки таких технологий одной из важнейших задач является составление алгоритма ввода структурных формул органических соединений. Задачу обычно решают путем кодирования химических структур с помощью специальных информационных языков — **линейных нотаций (представлений)**, позволяющих однозначно представлять

двухмерную структурную формулу соединения набором строк символов, удобных для ввода в память ЭВМ.

Можно выделить четыре основных типа химических нотаций:

1. Уникальные и однозначные:

одно соединение \longleftrightarrow одна нотация

А. *Традиционные, общепринятые*: «систематические» имена, которые приняты в рамках правил номенклатуры. Сюда относятся некоторые молекулярные формулы, например, CH_4 .

Б. *Нетрадиционные*: номенклатура ИЮПАК и линейная нотация Висвессера.

2. Уникальные и неоднозначные:

соединение 1 \searrow
соединение 1 \nearrow → одна нотация

А. *Традиционные, общепринятые*: большая часть молекулярных формул, например, C_4H_{10} .

Б. *Нетрадиционные*: любые фрагментарные коды (СВСС) или любые классификационные коды (такой, как Визелогле) в рамках своих правил.

3. Неуникальные и однозначные:

одно соединение \longleftrightarrow нотация 1
 \longleftrightarrow нотация 2

А. *Традиционные, общепринятые*: «систематические» имена, которые не ограничиваются рамками правил номенклатуры. Структурные формулы.

Б. *Нетрадиционные*: топологические коды, такие как Нортон-Оплер, Вальдо и Гайштагер.

4. Неуникальные и неоднозначные:

соединение 1 \longleftarrow нотация А \longrightarrow соединение 2
 \longleftarrow нотация Б \longrightarrow

А. *Традиционные, общепринятые*: основные тривиальные имена, такие как соли, масла, спирты и др.

Б. *Нетрадиционные*: любые фрагментационные или классификационные коды, которые не ограничиваются рамками своих правил.

В настоящее время существует множество систем кодирования и анализа химических структур. Под **системой кодирования** понимают совокупность лексических и грамматических правил, в которой каждой структурной формуле соответствует слово (цифра) в заданном алфавите (системе исчисления). Можно выделить три основных типа систем кодирования: **фрагментарные коды, линейные нотации и топологические методы**.

При фрагментарном представлении структур соединение рассматривается как сумма определенных фрагментов, каждый из которых имеет свое условное обозначение. Так, в частности, применяют линейную нотацию Висвессера, в которой используется 40 символов: 26 латинских заглавных букв, 10 цифр, три знака и пробел.

Буквами обозначаются структурные элементы (функциональные группы, бензольное ядро и т. д.) и места расположения заместителей в циклах, цифрами — длины углеродных цепей или размеры циклов. Символы всегда располагаются в порядке принятого в системе старшинства, что обеспечивает однозначность шифров.

Другим способом фрагментарного кодирования являются шифры ИЮПАК, в основе которых используется система Дайсона, базирующаяся на Женевской номенклатуре органических соединений. Например, запись формулы неопентил хлорида в системе Дайсона — C_3C_2Ch расшифровывается так: Ch — хлорид; C_3 — цепь из трех атомов углерода; C_2 — два С-атома (CH_3 -группы) у второго углеродного атома цепи. Шифр *n*-аминобензойной кислоты — $B6CX1N4$: B6 — бензольное кольцо; C — один алициклический С-атом; X1 — COOH-группа в положении 1; N4 — аминогруппа в положении 4.

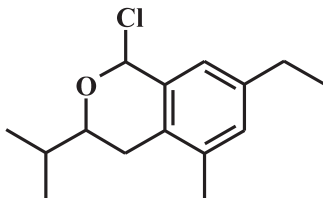
В отличие от систем Висвессера и Дайсона, дающих строго однозначный, т. е. канонический вид представления информации, существуют так называемые неканонические системы фрагментарного кодирования: структура одного и того же соединения может быть представлена различными кодами, которые с помощью компьютера переводятся в канонический вид. Такая система, в частности,

разработана и функционирует в ВИНТИ (Всероссийский институт научной и технической информации) для ввода структур органических соединений на магнитные носители. В основе «произвольно-блочной» системы кодирования ВИНТИ используется предварительная обработка структурной формулы — она расчленяется на циклические фрагменты, имеющиеся в них заместители и фрагменты-связи.

Способы фрагментарного кодирования чаще всего используются при создании различных автоматизированных библиотечных информационных систем, к основным назначениям которых следует отнести быстрый анализ и регистрацию новых данных, поиск и выдачу информации по запросу потребителя при компактном размещении большого массива данных на магнитных носителях. В то же время при фрагментарном представлении структур теряется ценная информация об общем строении молекулы, что является существенным ограничением их применения при решении различных аналитических задач по поиску корреляционных зависимостей типа «структура — свойство».

В основе представления химических структур с помощью линейных нотаций используются различные лингвистические модели, базирующиеся на преобразовании исходных структур в набор дескрипторов, характеризующих наличие или, наоборот, отсутствие определенных структурных фрагментов, а также физико-химические свойства отдельных структурных фрагментов или полностью молекулы как единого целого. Многие лингвистические модели построены на применении одного из четырех типов формальных грамматик Хомского. Наиболее часто в этих моделях используются следующие положения: вводятся определенные обозначения основных функциональных групп (линейные цепочки отображаются в порядке расположения в структурной формуле атомов или функциональных групп, боковые ответвления заключаются в круглые скобки); в сложных циклических системах вводится индексация точек начала, окончания и сопряжения циклов. Используемая для анализа представления химических структур грамматика включает в себя различные множества кодов атомов и функциональных групп, а также правила грамматики, описывающие определенный

класс органических соединений. Например, для циклической структуры 1-хлор-3-изопропил-5-метил-7-этил-2-окса-1,2,3,4-тетрагидронафталин предложена следующая линейная нотация кодирования:



C3C1(C3)#C10C1(G)C.1UC.2C2&
#C1UC(C2C3)C1UC(C3)C.2UC.1&

В этой записи использована система условных обозначений: C, O, G — атомы углерода, кислорода, хлора; () — боковые ответвления; # и & — начало и конец цикла; U — двойная связь; C1 и C2 — индексация точек сопряжения двух циклов.

В отличие от фрагментарного кодирования, форма представления линейных нотаций содержит значительно больше информации о структуре конкретного соединения, и тем не менее обработка данных в таком виде в процессе компьютерного структурного анализа в ряде случаев не позволяет решать, например, такие задачи, как однозначное установление изоморфизма структур, определение общих частей структур, оценка положения этих частей в структуре. Кроме того, в линейных нотациях химических структур часто используются неканонические системы кодирования.

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

которые в дальнейшем были использованы для конструирования новых активных соединений.

1.3. ПРЕДСТАВЛЕНИЕ ОРГАНИЧЕСКИХ МОЛЕКУЛ В КОМПЬЮТЕРНОМ ВИДЕ

Существует два основных типа представления молекул в компьютерном виде: линейное и табличное.

В свою очередь, линейное представление молекул включает в себя основные наиболее распространенные кодировки:

- 1) номенклатура (тривиальная или ИЮПАК);
- 2) линейная нотация Висвессера (WLN);
- 3) нотация ROSDAL;
- 4) представление SMILES;
- 5) нотация SLN;
- 6) единая система InChI, принятая ИЮПАК.

Табличные представления химических молекул также можно разделить на представления в виде:

- графов, матриц;
- продолжающихся таблиц.

2. ЛИНЕЙНЫЕ НОТАЦИИ

Граматики первых линейных химических нотаций представляли собой варианты адаптированных правил номенклатуры органических соединений (старшинство групп, выбор главной цепи, начало ее нумерации и т. п.). Номенклатура органических соединений изучается в курсе органической химии, поэтому здесь нет надобности рассматривать данный тип линейной кодировки молекул.

2.1. ЛИНЕЙНАЯ НОТАЦИЯ ВИСВЕССЕРА (WLN)

В 1948 г. Вильгельмом Висвессером (W. Wiswesser) был получен патент на линейное представление молекул для компьютерного анализа. Это был первый пример использования компьютеров для записи химических соединений и поиска среди них биологически активных соединений. Данная нотация широко использовалась вплоть до конца XX в.

Следует отметить, что линейная нотация Висвессера (Wiswesser Line Notation, WLN) относится к однозначным языкам, которые позволяют полностью воспроизводить структурную формулу соединения по ее линейной записи, в связи с чем они использовались, например, в информационно-поисковой системе CAS с 1953 г.

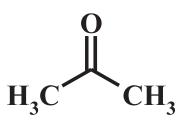
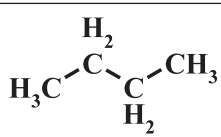
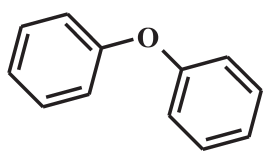
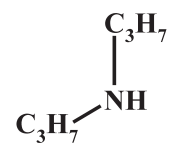
В словаре нотации Висвессера буквенные обозначения присвоены всем химическим элементам (с учетом валентности их атомов и ближайшего окружения), а также основным типам связей, циклов и функциональных групп.

Для записи в WLN используют 40 символов:

- буквы от «**A**» до «**Z**»,
- числа от «**0**» до «**9**»,
- символы «**&**», «**/**», «**-**» и **пробел**.

В табл. 4 приведены примеры записи органических соединений в WLN.

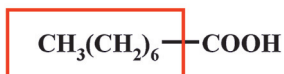
Примеры записи в WLN

Структурная формула	Запись в формате WLN	Пояснения
	1V1	Группа H ₃ C- (1) Группа -CO- (V)
	4	Группа CH ₃ CH ₂ CH ₂ CH ₃ (4)
	ROR	Группа фенил (R) Кислородный мостик -O- (O)
	3M3	Группа CH ₃ CH ₂ CH ₂ - (3) Группа -NH- (M)

Правила WLN

1. Углерод обозначается **числом**, X, Y или C:

— неразветвленные алканы обозначаются **числом**.



7



2

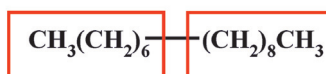
4



3



4



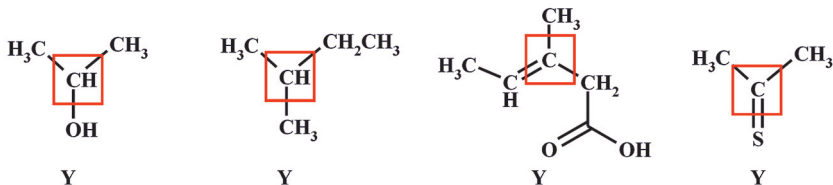
7



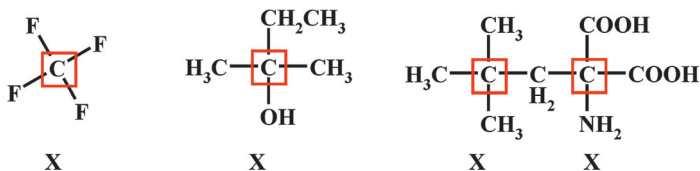
9

Линейные алканы (алкильные заместители), состоящие более чем из девяти атомов углерода, можно разбить на несколько меньших или перед числом поставить тире. Так, например, гексадекан ($C_{16}H_{34}$) можно обозначить как 7-9, либо как -16;

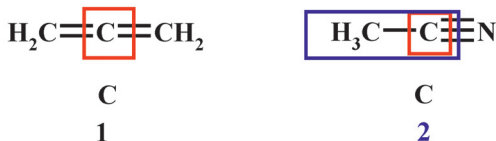
— углерод, соединенный с тремя тяжелыми атомами, обозначается Y:



— углерод, соединенный с четырьмя тяжелыми атомами, обозначается X:

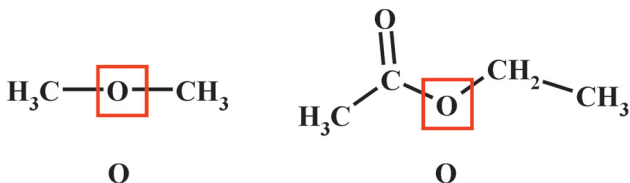


— символ C используется достаточно редко, только для углерода, не имеющего водородов и соединенного с двумя тяжелыми атомами, например, в аллене или нитриле (можно заменять числом):

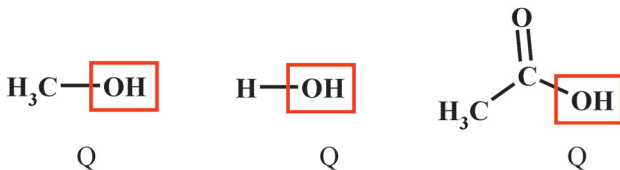


2. Кислород обозначается O, Q, V или W:

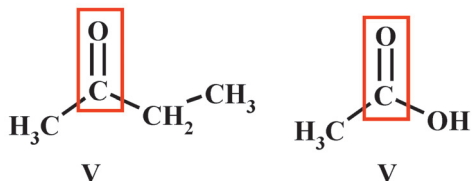
— кислород, соединенный с двумя тяжелыми атомами, например в эфирах, обозначается O:



— гидрокси-группа (OH) обозначается как Q:



— карбонильная группа (C=O) обозначается V:



— два кислорода у одного атома обозначаются как W, например NO₂ и SO₂ — как NW, SW:



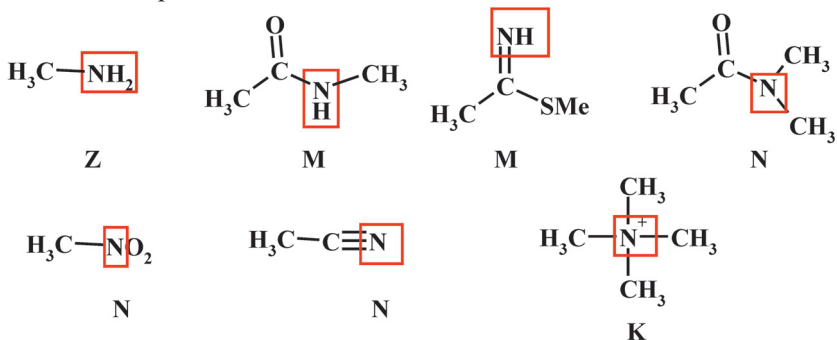
3. Азот обозначается K, M, N или Z:

— азот, соединенный с тремя тяжелыми атомами, обозначается N (amiNe);

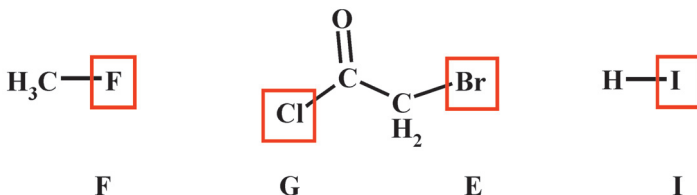
— буквой M обозначают группы -NH- или =NH (aMine);

— концевую аминогруппу (-NH₂) обозначают Z (последняя буква в алфавите);

— четвертичный атом азота обозначают K:

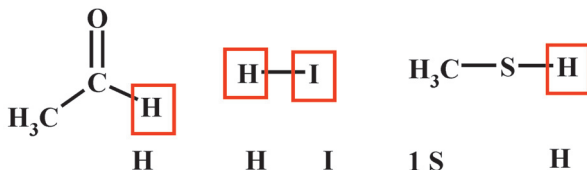


4. Другие гетероатомы обозначаются так: фтор как **F**, йод как **I**, хлор как **G**, бром как **E**, сера обозначается как **S**, фосфор обозначается как **P**:

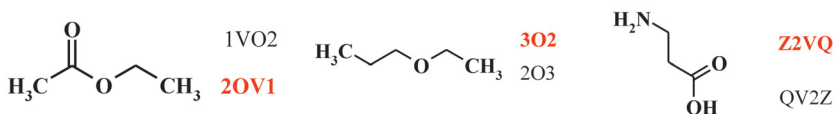


5. Двойные связи обозначаются **U**, тройные — **UU**.

6. Водород никак не обозначается, только в исключительных случаях обозначается как **H**:

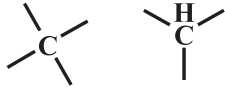


В табл. 5 приведены основные обозначения функциональных групп по системе WLN. Запись соединения в системе WLN можно осуществить по-разному, в зависимости от того, с какого края молекулы начинать кодировать структуру:

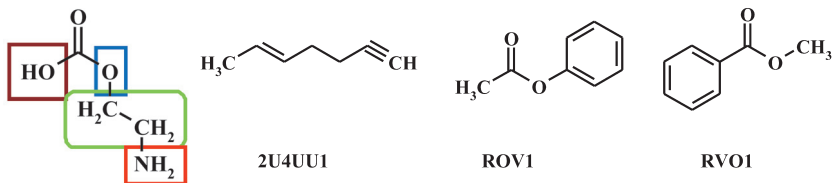


Принято записывать по алфавиту (обратный алфавит, т. е. **Z > A**), причем цифра младше буквы, цифры располагаются по старшинству (2 больше, чем 1). Таким образом, правильная запись этилацетата — **2OV1**. Для этоксипропана EtOPr — **3O2**, а не 2O3, для β-аланина ($\text{H}_2\text{NCH}_2\text{CH}_2\text{COOH}$) — **Z2VQ**, а не QV2Z.

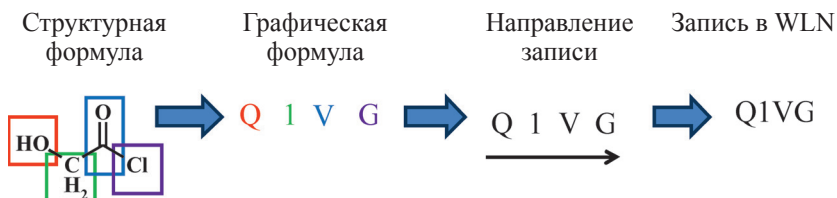
Обозначение функциональных групп в системе WLN

Название	Формула	WLN
Водород	H	H
Алкан, алкил	$C_n H_{2n+2}$, $-C_n H_{2n+1}$	Число n
Алкен	$-CH=CH-$	U
Алкин	$-C\equiv C-$	UU
Аллен	$-CH=C=CH-$	U1U UCU
Разветвление		X, Y
Галогены	F, Cl, Br, I	F, G, E, I
Амины	NH_2 -NH- $-N=$ $=N^+=$	Z M N K
Спирты	-OH	Q
Эфиры	-O-	O
Диоксогруппа	NO_2	NW
	SO_2	SW
Кетоны	$-C(=O)-$	V
Альдегиды	-CHO	VH
Кислоты	-COOH	VQ
Сложные эфиры	-COO-	VO
Нитрилы	$-C\equiv N$	1UUN CUUN
Имины	$-C=NH$	CUM

Примеры:

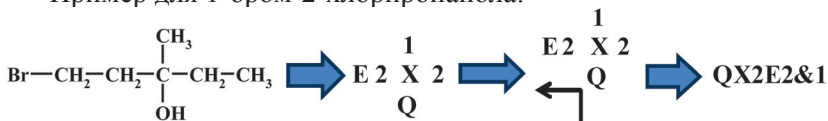


Разберем пример **кодирования** в системе WLN. Первый шаг, после того как записана структурная формула, это поиск функциональных групп и запись их в виде графической формулы. Вторым шагом — выбор направления записи по правилу «старшинства» (группы располагаем по алфавиту и старшинству — обратный алфавит и бóльшая цифра). Для молекулы гидроксиацетилхлорида получилась следующая запись в формате WLN:

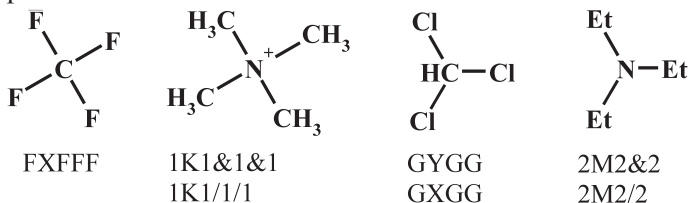


При наличии разветвлений в молекуле, после обозначения атома с несколькими связями, например, N, X, K или Y, запись продолжают по правилам старшинства. Конец разветвления обозначают «&» или «/», в случае если ответвление оканчивается на терминальную группу, например, Q, H, G, I, F, NW, знак окончания можно опустить и после этого вернуться к началу разветвления и записать следующую цепочку. Необходимо помнить, что знаки **X** и **K** имеют по два окончания **& (/)**, знаки **Y** и **N** имеют по одному — **& (/)**. Если в цепочке есть еще разветвления, то после окончания цепочки возвращаются к **ближайшему**.

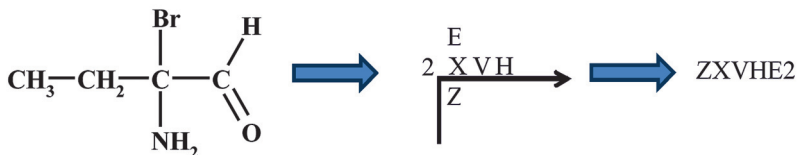
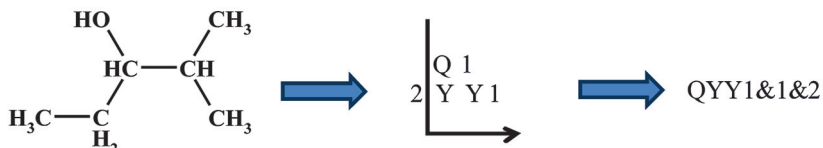
Пример для 1-бром-2-хлорпропанола:



Примеры для тетрафторметана, тетраметиламмония, хлороформа и триэтиламина:



Примеры алгоритма кодирования в WLN:

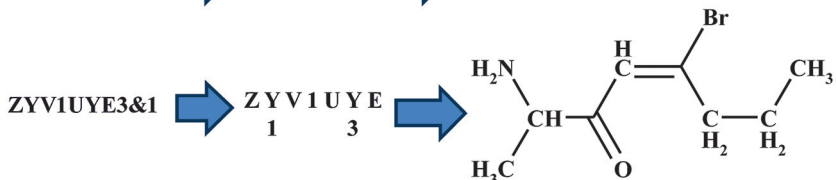


Рассмотрим алгоритм **декодирования** записи в WLN.

Шаг первый: разбивают коды функциональных групп по цепочкам, учитывая разветвления, и записывают коды соответственно разветвлениям.

Шаг второй: декодируют функциональные группы, расставляя необходимые связи между атомами. Ниже представлены примеры декодирования записи в WLN.

Запись в WLN Графическая формула Структурная формула



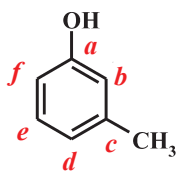
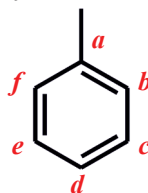
Правила записи циклических соединений в WLN

1. Ароматические моноциклические заместители (фенильные) обозначают буквой **R**.

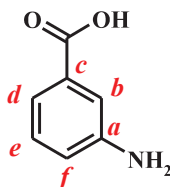
2. Карбоциклические системы (моно и конденсированные) обозначаются буквами: начало цикла как **L**, конец цикла — **J**, между этими буквами располагают информацию о размере цикла и двойных связях. Размер цикла обозначают **цифрой** сразу после буквы **L**.

3. Запись заместителей в бензольном кольце осуществляется следующим образом: первый заместитель (или соединение с цепочкой атомов) помечают как позицию **a**, затем по алфавиту.

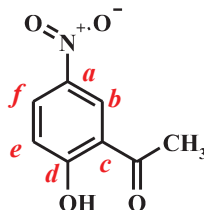
Запись ведут соответствующей заглавной буквой через пробел, после окончания цикла (буквы **J** или **R**).



QR C1



ZR CVQ



WNR CV1 DQ

4. Если цикл полностью гидрированный — добавляют букву **T** перед **J**.



L3TJ



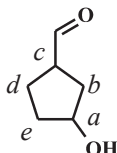
L4TJ



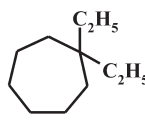
L4J



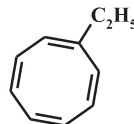
L6TJ



L5TJ AQ CVH



L7TJ A2 A2



L8J A2

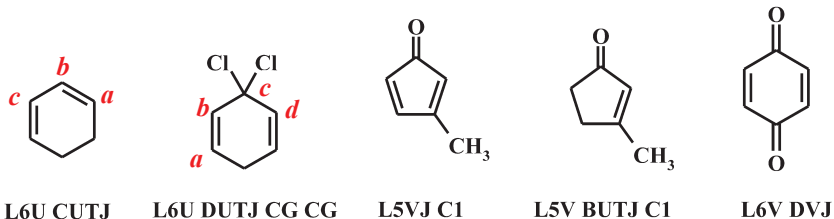
Если цикл гидрирован частично, то в случае, когда только **один атом углерода насыщен**, его положение обозначают символом **H** между буквами **L(T)** и **J**.

Если в цикле **более одного насыщенного атома**, отмечают положение двойных связей и добавляют букву **T**.



L5 AHJ

В случае наличия в цикле **карбонильной функции**, ее обозначают буквой **V** после указания положения.



5. Гетероциклы записывают с буквы **T** (начало цикла), числом записывают размер цикла, затем отмечают гетероатомы и их положение, конец цикла обозначается буквой **J**. Примеры записи гетероциклов и заместителей приведены на рис. 1.

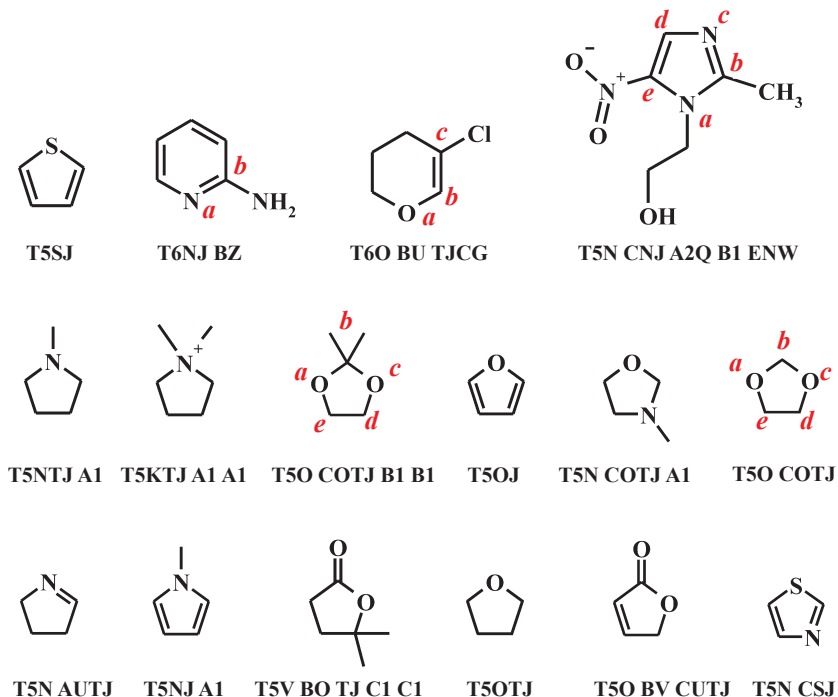
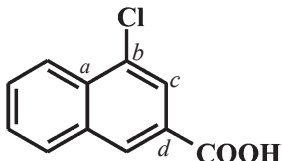
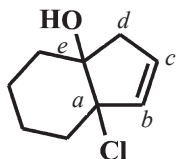


Рис. 1. Примеры записи гетероциклических соединений

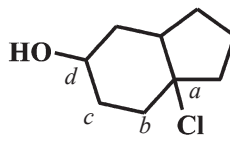
6. Би- или полициклические соединения записывают, начиная с буквы **L** или **T**, в зависимости от того, входит ли в состав конденсированного соединения гетероцикл или нет, следующие две цифры обозначают размер сочлененных циклов и окончание цикла **J**, положения замещения начинают нумеровать с атома сочленения так, чтобы заместители имели наименьшую букву (рис. 2).



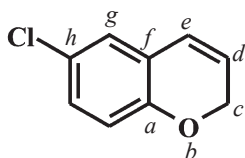
L66J BG DVQ



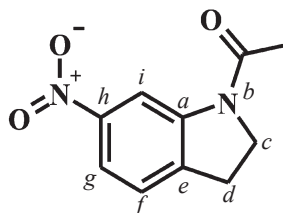
L56 BUJ AG EQ



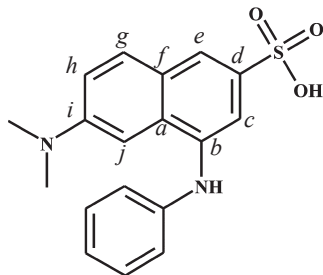
L65TJ AG DQ



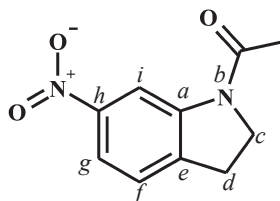
T66 VO CHJ HG



T56 BNT&J BV1 HNW



L66J BMR & DSWQ IN1&1



T56 BN T&J BV1 HNW

Рис. 2. Примеры записи бициклических соединений

При записи в линейном виде полициклических соединений выбирают первые два конденсированных цикла (например **65**), нумеруют буквами по кругу, третий цикл записывают с буквой соединения перед первыми двумя цифрами (рис. 3).

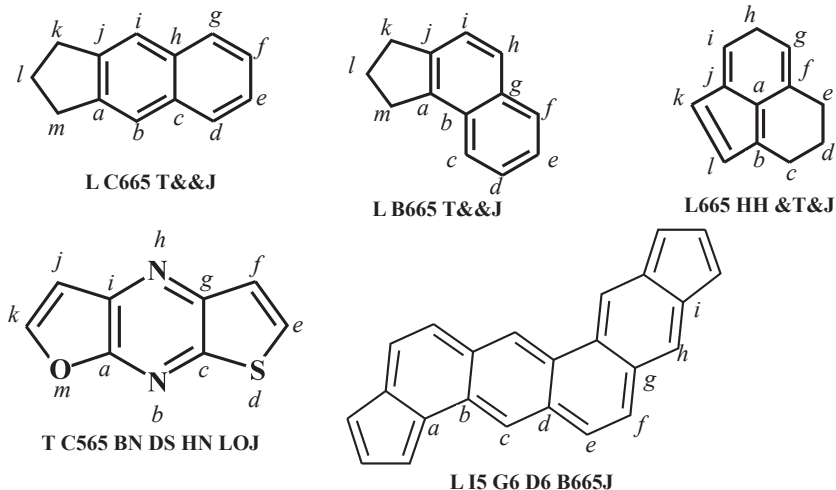


Рис. 3. Примеры записи полициклических соединений

Несмотря на то что система WLN практически не встречается сейчас в научной литературе, достаточно много компьютерных программ по QSAR используют данную систему записи для алгоритма поиска фармакофорных групп.

2.2. ПРЕДСТАВЛЕНИЕ ОРГАНИЧЕСКИХ СТРУКТУР В ФОРМАТЕ ROSDAL

Кодировка ROSDAL (Representation of Organic Structure Description Arranged Linearly) была создана в 1985 г. в Институте Бельштейна для перевода справочника Бельштейна в компьютерную базу данных Beilstein Institute. Применяется для названий соединений в программе ISISDraw и в базе данных Reaxys.

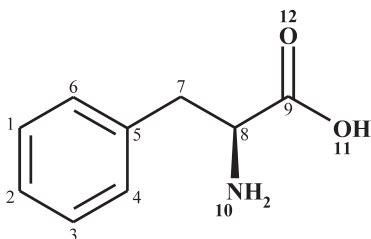
Правила ROSDAL

1. Вся структура нумеруется, символы атомов пишутся сразу после номера (символ углерода С не пишется), водороды не обозначаются или нумеруются наравне со всеми атомами;

2. Одинарные связи обозначаются «-», двойные — «=», тройные — «#», другие — «?»;

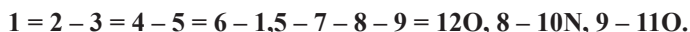
3. Разветвления перечисляются через запятые.

Рассмотрим алгоритм кодирования в формате ROSDAL на примере молекулы фенилаланина.

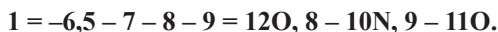


Фенилаланин

После того как структура пронумерована, начинаем запись кода молекулы с учетом кратности связи. Сначала записываем бензольное кольцо, начиная с углерода под номером 1, затем через запятую все оставшиеся звенья молекулы:



Существует также второй вариант записи:



2.3. СПЕЦИФИКАЦИЯ УПРОЩЕННОГО ПРЕДСТАВЛЕНИЯ МОЛЕКУЛ В СТРОКЕ ВВОДА (SMILES)

В настоящее время наиболее распространенным видом линейных нотаций является SMILES (Simplified Molecular Input Line Entry System — спецификация упрощенного представления молекул в строке ввода).

Первоначальный вариант кодировки был разработан Артуром Вейнингером (Arthur Weininger) и Дэвидом Вейнингером (David Weininger) в 1986 г. Система SMILES относится к нотации однозначного описания состава и структуры молекулы химического вещества

с использованием строки символов ASCII. Название в английском языке является омонимом к слову *smiles* (улыбки), однако пишется только заглавными буквами. Строка символов, составленная по правилам SMILES, может быть преобразована многими молекулярными редакторами в двумерную или трехмерную структурную формулу молекулы.

Правила SMILES

1. **Атомы.** Все неводородные атомы пишутся своими символами в квадратных скобках (первая буква заглавная, вторая строчная). Для элементов органической химии, таких как B, C, N, O, P, S, F, Br и Cl, разрешается запись без квадратных скобок, при этом также можно не указывать присоединенные к этим элементам атомы водорода, количество которых рассчитывается по минимальной нормальной валентности и указанным связям. Принимается следующая минимальная нормальная валентность: B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6), H (1). Атомы в ароматических кольцах указываются строчными буквами, например, алифатический углерод обозначают заглавной буквой «C», ароматический — строчной буквой «c». Атомы с валентностью, отличающейся от минимальной нормальной, записывают в квадратных скобках, как и остальные элементы (табл. 6).

Таблица 6

Обозначение атомов в SMILES

Обозначение в SMILES	Название вещества	Формула
C	Метан	CH ₄
P	Фосфин	PH ₃
N	Аммиак	NH ₃
S	Сероводород	H ₂ S
O	Вода	H ₂ O
Cl	Хлороводород	HCl
[S]	Элементарная сера	S
[Au]	Золото	Au

В случае наличия заряда атом записывают в скобках с обязательным указанием добавленных водородов и заряда. Количество атомов водорода указывают числом после символа H. Формальный заряд указывают символами «+» или «-» после символов водорода (табл. 7). Запись формы [Fe⁺⁺⁺] является эквивалентной записи [F+3].

Таблица 7

Обозначение зарядов в SMILES

Обозначение в SMILES	Название вещества
[H+]	Протон
[Fe+2]	Катион железа (II)
[OH-]	Гидроксид-анион
[Fe+++]	Катион железа (III)
[OH3+]	Гидроксоний катион
[NH4+]	Катион аммония

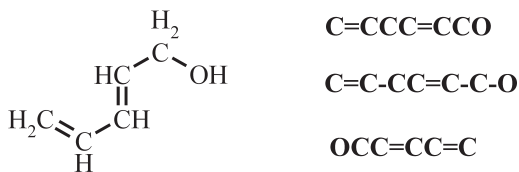
2. **Связи.** Одинарные, двойные и тройные обозначаются «-», «=» и «#» соответственно. Подразумевается, что записанные вместе атомы соединены между собой одинарной и ароматической связью, т. е. эти связи можно не указывать (табл. 8).

Таблица 8

Обозначение связей в SMILES

Обозначение в SMILES	Название вещества	Формула
CC	Этан	CH ₃ CH ₃
C=O	Формальдегид	CH ₂ O
C=C	Этен	CH ₂ =CH ₂
C#C	Ацетилен	HC≡CH
O=C=O	Диоксид углерода	CO ₂
O=CO	Муравьиная кислота	HCO ₂ H
COC	Диметиловый эфир	CH ₃ OCH ₃
C#N	Цианистая кислота	HCN
CCO	Этанол	CH ₃ CH ₂ OH
[H][H]	Молекулярный водород	H ₂

Для линейных молекул запись в SMILES возможно осуществить несколькими способами. Так, например, для 6-гидрокси-1,4-гексадиена возможны три эквивалентные записи в SMILES:



Ионные связи отмечают как отдельные молекулы через точку:
[Na+].[Cl-] — хлорид натрия (NaCl);
CC[O-].[Na+] — этилат натрия (C2H5ONa).

3. Разветвления. Боковые цепи отделяют от основной цепи круглыми скобками. Например, пропионовая кислота записывается как CCC(=O)O. Форма записи трифторметана выглядит как C(F)(F)F, однако такая запись неудобна для чтения из-за своей перегруженности скобками, поэтому ту же молекулу можно записать в неканонической форме как FC(F)F (табл. 9).

Таблица 9

Примеры записи разветвленных соединений в SMILES

<chem>CCN(CC)CC</chem>	<chem>CC(C)C(=O)O</chem>	<chem>C=CC(CCC)C(C(C)C)</chem>
Триэтиламин	Изомасляная кислота	3-пропил-4-изопропил-гептен-1

4. Циклические соединения. Циклические структуры записывают с разрывом одной из связей. Эти «разорванные» связи нумеруются в произвольном порядке. Атомы, находящиеся на концах

разорванной при построении основной линии связи, обозначаются одним и тем же номером. Циклические атомы записываются прописными буквами. Например, запись циклогексана выглядит следующим образом: C1CCCCC1 (рис. 4).

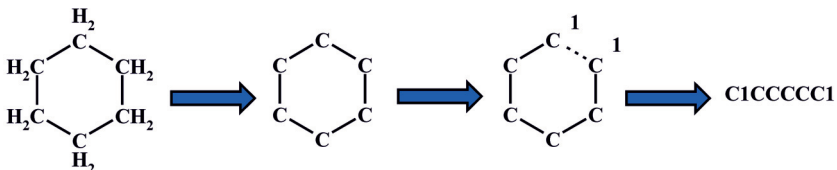


Рис. 4. Запись циклогексана в нотации SMILES

В зависимости от того, какую линию связей выбрать за основную, может существовать несколько вариантов записи SMILES-кода, при этом они имеют равное право на свое существование. Запись 1-метил-3-бромциклогексена-1 представлена на рис. 5.

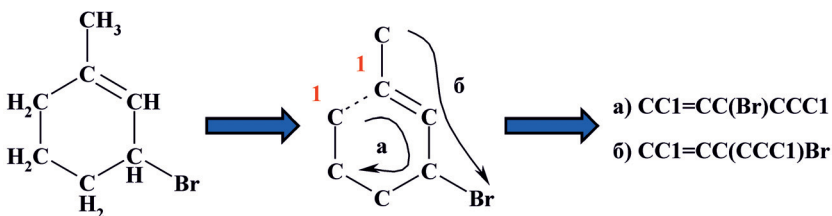


Рис. 5. Запись 1-метил-3-бромциклогексена-1 в нотации SMILES

Один атом может одновременно входить в состав нескольких циклов. Так, в записи кубана один и тот же атом углерода может попасть на «вход» сразу в нескольких циклах: C12C3C4C1C5C4C3C25. Соответствующая запись представлена на рис. 6.

Пример записи для более сложной молекулы — морфина представлен на рис. 7.

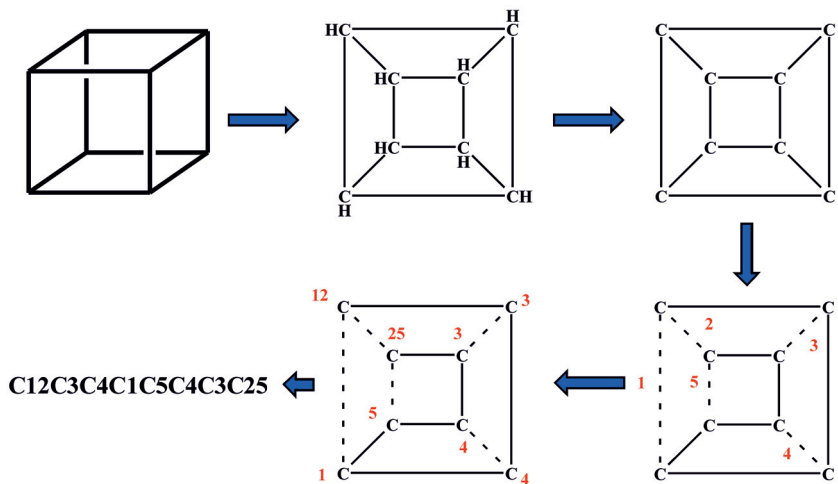


Рис. 6. Запись кубана в нотации SMILES

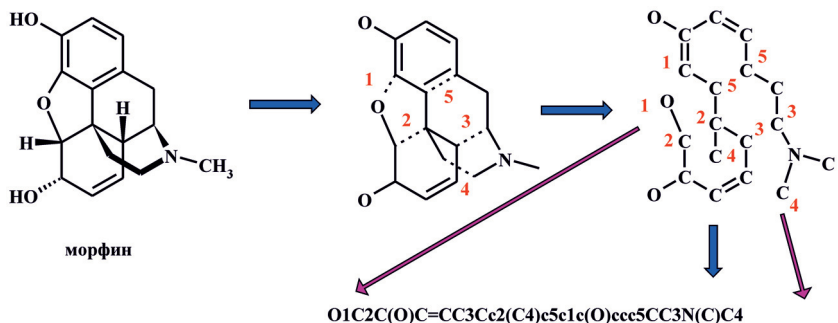


Рис. 7. Запись морфина в нотации SMILES

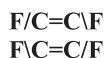
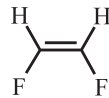
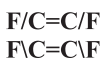
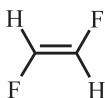
Если в структуре присутствуют 10 или более циклов, то перед двузначным номером цикла необходимо ставить знак процента (%). Например, запись C2%13%24 означает, что у атома углерода есть три разрыва циклов — 2, 13 и 24.

5. **Изотопы** записывают перед символом атома, в этом случае квадратные скобки обязательны (табл. 10).

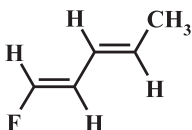
Обозначение зарядов в SMILES

Обозначение в SMILES	Название вещества
[12C]	Углерод-12
[13C]	Углерод-13
[C]	Углерод (изотоп не указан)
[13CH4]	С-13 метан
[236U]	Уран-238
[15NH4+]	Катион аммония-15

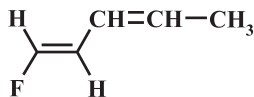
6. Конфигурация относительно двойной связи. Цис-транс-изомерия относительно двойной связи обозначается символами «/» или «\» в зависимости от направления связи, эти символы используются вместо обозначения одинарной или ароматической связи. Они показывают относительное расположение соединенных атомов и имеют смысл только в случае обозначения обоих атомов, которые присоединены к двойной связи. Например, правильное обозначения E- и Z-изомеров 1,2-дифторэтанов следующее:



SMILES также позволяет отмечать случаи, если известна конфигурация только одной из нескольких двойных связей. Данное положение может быть продемонстрировано следующим примером:

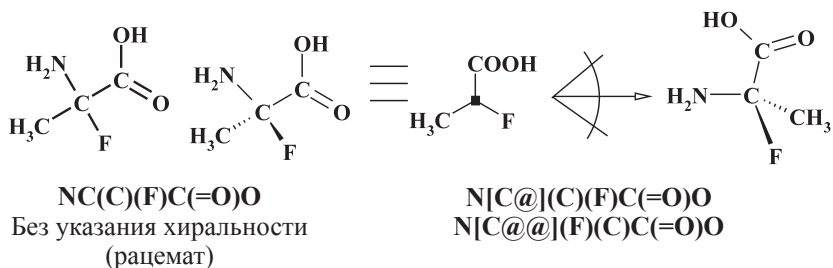


Известна конфигурация
двух связей



Известна конфигурация
только первой связи

7. **Конфигурация асимметрического центра.** Система SMILES позволяет обозначать хиральный атом, т. е. энантимеры записываются разными способами. *Стереизомерия тетраэдрического атома* обозначают символами «@» и «@@». Если хиральность отсутствует, то у данного атома ничего дополнительного не записывают. Запись конфигурации хирального центра осуществляют согласно визуальному мнемоническому правилу: смотрят со стороны первого атома на хиральный порядок из записи молекулы в SMILES, если три атома располагаются против часовой стрелки, то записывают у хирального атома конфигурацию @, если атомы располагаются по часовой стрелке — @@. Таким образом, в случае записи @ следующий атом располагается слева от смотрящего, второй — справа, третий — вверху. Записи N[C@](C)(F)C(=O)O и N[C@@](F)(C)C(=O)O являются эквивалентными.



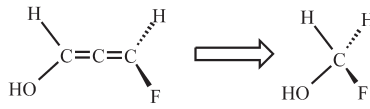
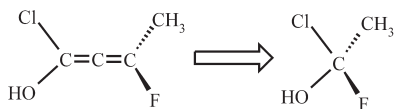
Если хиральный атом соединен только с тремя тяжелыми атомами, т. е. один из четырех возможных заместителей — водород, то в таком случае в квадратных скобках указывают либо атом водорода как заместитель, либо углерод как «неорганический» элемент, с обязательным добавлением водорода, у которого показывают хиральную конфигурацию символом @ (табл. 11). При этом запись хиральности определяется *порядком записи атомов* в SMILES.

Стереизомерия алленового типа. Стереизомерия алленов, содержащих четное количество атомов при двойных связях, соответствует цис-транс-изомерии и, соответственно, записывается символами \ и /. Асимметрично замещенные аллены, содержащие нечетное количество атомов при двойных связях, обладают аксиальной хиральностью (точечная группа D_{2d}) — хиральным центром

Варианты записи энантимеров аланина в SMILES

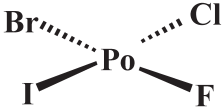
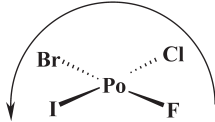
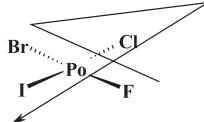
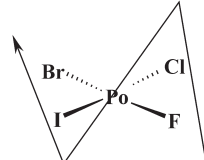
Структура изомера аланина	Запись в SMILES
	<chem>N[C@@]([H])(C)C(=O)O</chem> <chem>N[C@@H](C)C(=O)O</chem> <chem>N[C@H](C(=O)O)C</chem> <chem>[H][C@](N)(C)C(=O)O</chem> <chem>[C@H](N)(C)C(=O)O</chem>
	<chem>N[C@]([H])(C)C(=O)O</chem> <chem>N[C@H](C)C(=O)O</chem> <chem>N[C@@H](C(=O)O)C</chem> <chem>[H][C@@](N)(C)C(=O)O</chem> <chem>[C@@H](N)(C)C(=O)O</chem>

является центральным атом с двумя двойными связями. В данном случае хиральность обозначается как @AL1 (или @) и @AL2 (или @@). Правила записи тетраэдрического центра идентичны, данный центр можно «получить», мысленно совмещая концевые атомы алленовой системы, как показано ниже:

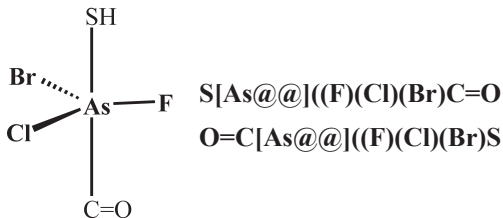


Планарная хиральность обозначается как SP (square-planar). Возможные варианты записи — @SP1, @SP2, @SP3 представлены в табл. 12.

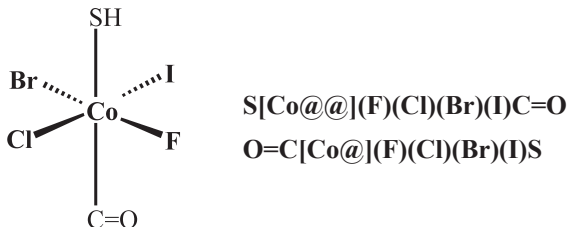
Варианты записи планарной хиральности в SMILES

<p>Запись в SMILES для</p> 	<p>Принцип записи</p>
<p><chem>F[Po@SP1](Cl)(Br)I</chem></p>	<p>SP1 — подковообразная (U-образная) запись заместителей</p> 
<p><chem>F[Po@SP2](Br)(Cl)I</chem></p>	<p>SP2 — запись заместителей 4-образным способом</p> 
<p><chem>F[Po@SP3](Cl)(I)Br</chem></p>	<p>SP3 — Z-образная запись заместителей</p> 

Тригональная-бипирамидальная хиральность обозначается как ТВ (trigonal-bipyramidal). Возможны значения от @TB1 до @TB20. Знак @TB1 (или @) используется в случае расположения первого и последнего заместителя по вертикальной оси, а оставшиеся три заместителя располагаются против часовой стрелки в плоскости перпендикулярной оси. Знак @TB2 (или @@) используется в случае расположения трех заместителей по часовой стрелке, и т. д.



Октаэдральная хиральность обозначается как ОН. Возможны значения от @ОН1 до @ОН30. Знак @ОН1 (или @) используется в случае расположения первого и последнего заместителя по вертикальной оси, а оставшиеся четыре заместителя располагаются против часовой стрелки в плоскости перпендикулярной оси.



Особенности описания органических молекул в SMILES

Ароматические соединения. Запись ароматических связей в SMILES, как отмечалось выше, фиксируется знаками «:» или строчными буквами. Например, корректная запись бензола представляется двумя способами: c1ccccc1 или C1:C:C:C:C:C1. Несмотря на это, допустима запись C1=CC=CC=C1, хотя в данном случае подразумевается гипотетическая структура гексатриена. Это возможно, так как в системе SMILES не оговариваются длины связей. Определение ароматичности соединений включает в себя также и правило Хюккеля — $4n + 2$ π -электронов, соответственно, циклические соединения, содержащие $4n$ π -электронов, относятся к антиароматическим.

В SMILES антиароматические соединения записываются аналогично ароматическим, например, записи c1ccc1 и c1ccccccc1 будут соответствовать циклобутadienu и циклооктатетраену, которые более корректно записать как C1=CC=C1 и C1=CC=CC=CC=C1. Следует также отметить, что гетероциклические ароматические соединения записывают строчными буквами, включая и символы гетероатомов. Положение атома водорода у атома азота пиррольного типа записывают в квадратных скобках [nH]. Аналогичным образом записывают заряды у мезоионных соединений и у N-оксидов. Примеры представлены на рис. 8.

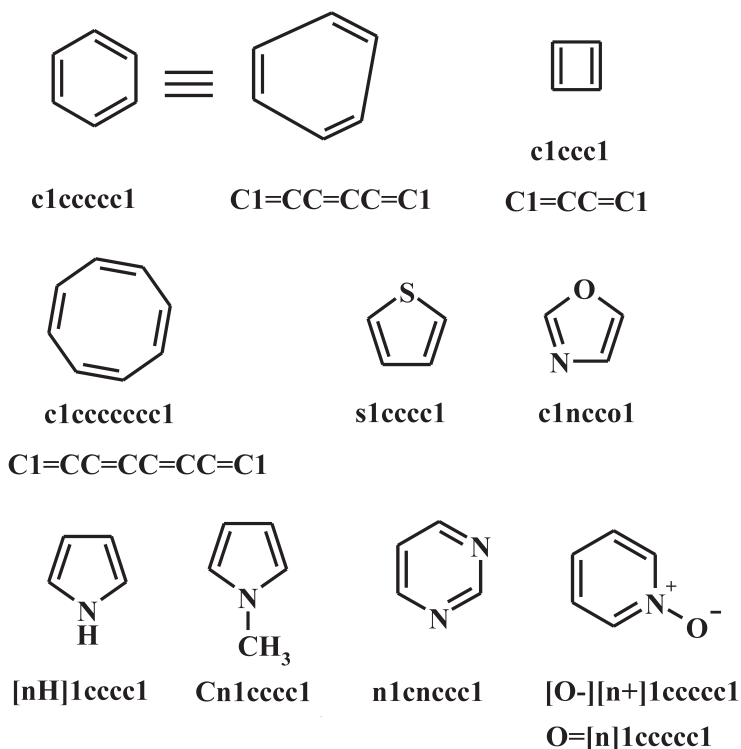


Рис. 8. Примеры записи антиароматических соединений в нотации SMILES

Связи. SMILES не регламентирует валентности атомов, которые записывают в молекулярную структуру. Фактически, используя SMILES, можно записать различные валентности в одной и той же структуре. Например, нитрометан может быть записан как CN(=O)=O или, используя формальные заряды и валентность, — как C[N+](=O)[O-]. Оба представления являются правильными. Обычно химики предпочитают записи с использованием формальных валентностей и зарядов. Например, запись диазометана C=[N+]=[N-] предпочтительней, чем C=[N]=[N].

Таутомеры в SMILES представляются как разные структуры. Не существует таких понятий, как «таутомерная связь», «мобильный

атом водорода» или «мобильный заряд». Пример изображения таутомерных форм представлен на рис. 9. Таутомерные формы одного и того же соединения различаются значительно.

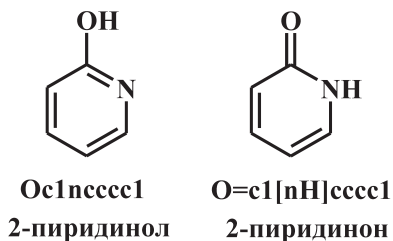


Рис. 9. Примеры записи таутомеров в нотации SMILES

2.4. ЛИНЕЙНАЯ НОТАЦИЯ СИБИЛ (SLN)

Данная кодировка (Sybyl Line Notation) была создана в 1986 г. Ее синтаксис базируется на правилах SMILES, позволяя при этом вставлять спецификации структур Маркуша, применяемые для патентирования химических структур (т. е. допускаются обозначения R и R_x в качестве боковой цепи, X — заместитель). Принципиальным отличием является то, что все атомы, включая атомы водорода, записываются, при этом атомы водорода ставятся после того атома, к которому присоединены. В квадратных скобках пишется только спецификация атома, например, изотопный состав C[I=14] — изотоп углерода-14. Ароматические связи можно обозначать только знаком «двоеточие», упрощение в виде написания строчными буквами не допускается.

Правила SLN

1. Пишутся все атомы (включая атомы водорода).
2. Одинарные связи не отмечаются, двойные обозначаются «=», тройные — «#».
3. Для указания разветвления используются скобки ().
4. Ароматические атомы обозначаются знаком «двоеточие».

5. Кольца нумеруются цифрами — цифра ставится у тех атомов, которые образуют кольцо.

6. Цис-транс-изомерия относительно двойной связи обозначается буквами «s» или «t».

7. Символ @ указывает на закрытие кольца.

В качестве иллюстрации данных правил можно привести следующие примеры:

CH₄

метан (CH₄)

NH₂

амин (-NH₂)

Na.OH

Na щелочь

HC(=O)OH

муравьиная кислота

CH₃C(=O)OH

уксусная кислота

C[1]H₂CH₂CH₂CH₂CH₂CH₂[@1]

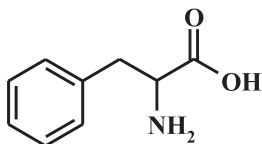
циклогексан

O[1]:CH:CH:CH:CH[:@1]

фуран

CH₃CH=[s=t]CHCH₃

транс-бутен



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

SLN поддерживается корпорацией *Tripos Inc* и используется в интеллектуальной системе CONCORD, а также для быстрого обмена информацией между различными стандартами записи. Ее правила записи очень интуитивны и их легко запомнить.

3. ТАБЛИЧНЫЕ ПРЕДСТАВЛЕНИЯ

3.1. ПРЕДСТАВЛЕНИЕ 3D-СТРУКТУР

3D-структура молекулы тесно связана с огромным разнообразием физических, химических и биологических свойств. Трехмерная структура молекулы называется **конформацией молекулы**. Следует отметить, что, как и в природе, существует иерархия в представлении химических соединений. Так, исходя из строения молекулы, можно узнать стереохимическую информацию, далее — конфигурацию молекулы (более детальное описание молекулы) и затем уже соответствующую 3D-структуру (рис. 10).

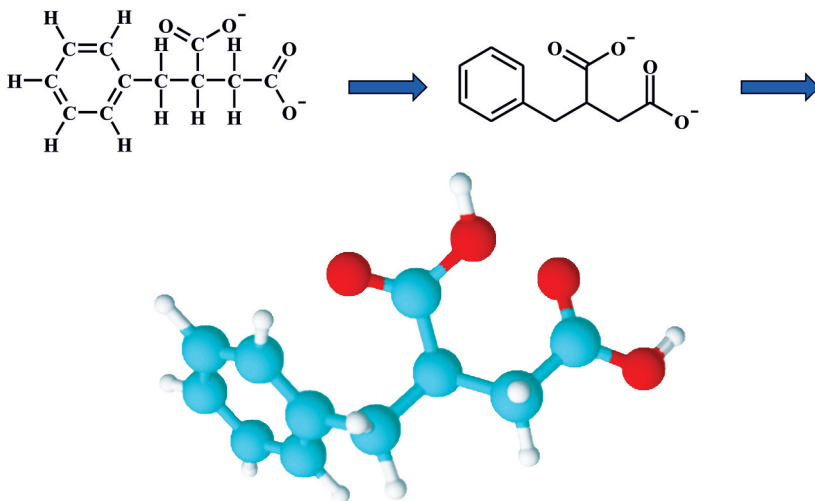
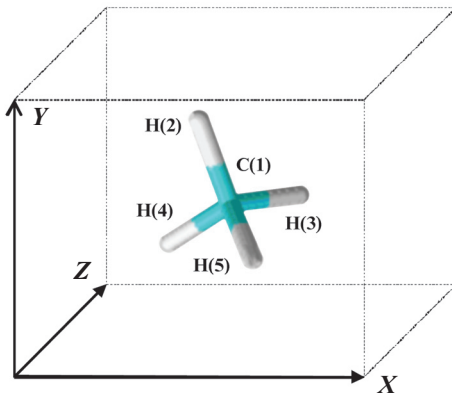


Рис. 10. От строения к конфигурации и затем к конформации (3D-структуре) молекулы на примере 2R-бензилсукцината

Для представления химической структуры в трехмерном пространстве в основном применяются два метода. Они используют различные координатные системы для описания пространственного

расположения атомов молекул. Изображение молекулы в декартовой системе координат представлено на примере метана (рис. 11).



	<i>X</i>	<i>y</i>	<i>Z</i>
C1	-0.0127	1.0858	0.0080
H2	0.0021	-0.0041	0.0020
H3	1.0099	1.4631	0.0003
H4	-0.5399	1.4469	-0.8751
H5	-0.5229	1.4373	0.9048

Рис. 11. Система декартовых координат и декартовы координаты метана

Наиболее широко используемым способом является выбор **декартовой системы координат**, т. е. кодирование *X*-, *Y*- и *Z*-координат каждого атома. Для каждого атома декартовы координаты можно перечислить в одном ряду. На рис. 11 дана иллюстрация этого метода на примере молекулы метана.

С помощью формата *XYZ*-координат можно точно установить геометрию молекулы, в которой количество атомов будет указано в декартовых координатах. Данный формат используется в программах компьютерной химии для импорта и экспорта геометрии. Эти единицы выражаются в большинстве случаев в ангстремах.

Число атомов будет указано в первой строке записи формата, комментарии — во второй строке, а далее находятся строки атомных координат.

Формат состоит из следующей записи:

```
<number of atoms>
comment line
atom_symbol1 x-coord1 y-coord1 z-coord1
atom_symbol2 x-coord2 y-coord1 z-coord2...
atom_symboln x-coordn y-coordn z-coordn
```

Файлы, использующие XYZ-формат, условно имеют расширение .XYZ.

Информацию о молекуле также можно получить с помощью XYZ-координат при дополнении таблицы новыми столбцами, где помимо координат можно увидеть связи одних атомов с другими. Рассмотрим это на примере молекулы этанола (рис. 12).

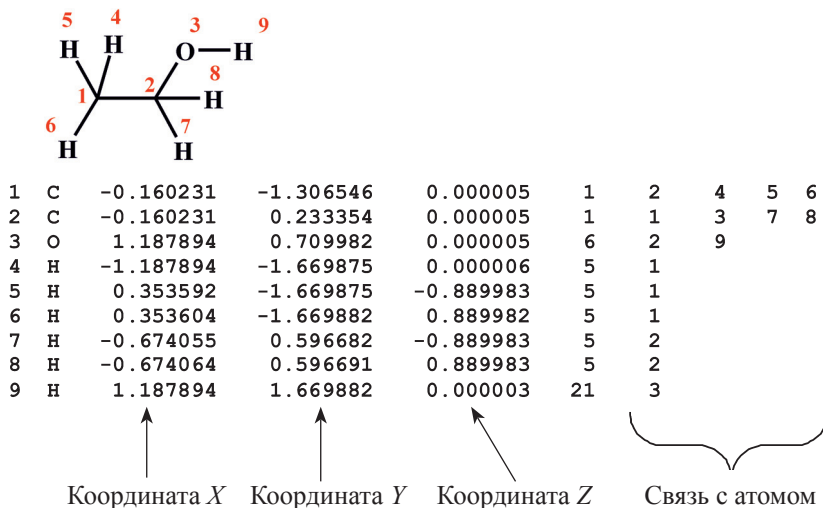


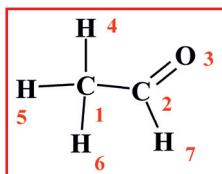
Рис. 12. Таблица с координатами и связями этанола

В большую часть стандартных форматов файлов, содержащих 3D-информацию о химической структуре, входят не только 3D-координаты атомов, но также и продолжающиеся таблицы.

3.2. ПРОДОЛЖАЮЩИЕСЯ ТАБЛИЦЫ (ТАБЛИЦЫ СОЕДИНЕНИЙ)

Продолжающиеся таблицы (Connection Table, СТ) — это такое представление молекулы, которое может быть получено с помощью перечисления в табличной форме атомов и связей молекулярной структуры. Таблица соединений — отображение состава вещества и связей между атомами в табличной форме. Необходимо различать каждый атом и каждую связь в молекуле. Это достигается при помощи перечисления атомов и связей и показа того, каким образом атомы связаны между собой.

Существует много способов представления таблицы соединений. Первый способ состоит в произвольной маркировке каждого атома молекулы и расположении их в списке атомов (рис. 13).



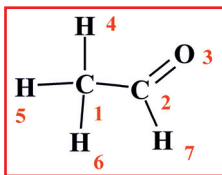
Список атомов	
1	С
2	С
3	О
4	Н
5	Н
6	Н
7	Н

Список связей		
1-й атом	2-й атом	Порядок связи
1	2	1
2	3	2
2	7	1
1	4	1
1	5	1
1	6	1

Рис. 13. Продолжающаяся таблица этаноля

Во второй таблице находится информация о связи с указанием атомов, которые соединяются этой связью. Кроме того, порядок связи соответствующего соединения записывается как целочисленный код (1 = одинарная связь, 2 = двойная связь и т. д.) и находится в третьей колонке.

В качестве альтернативы продолжающаяся таблица может быть представлена в форме одной избыточной СТ (рис. 14). Каждый атом упоминается дважды, сведения о водороде стандартны.

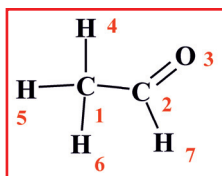


Атом- ный ин- декс	Эле- мент	1-й ин- декс атома	По- рядок связи	2-й ин- декс атома	По- рядок связи	3-й ин- декс атома	По- рядок связи	4-й ин- декс атома	По- рядок связи
1	С	2	1	4	1	5	1	6	1
2	С	1	1	3	2	7	1		
3	О	2	2						
4	Н	1	1						
5	Н	1	1						
6	Н	1	1						
7	Н	2	1						

Рис. 14. Избыточная продолжающаяся таблица этанала

На рис. 14 показано, что первые две колонки таблицы дают индекс атома и соответствующий символ элемента. Список связей интегрируется в табличную форму, в которой описаны атомы. Атом может быть связан с другими атомами: атом с индексом **1** связан с атомами **2, 4, 5 и 6**. Это можно записать в одну линию. Порядок связи обозначается **1** для одинарной связи, **2** — для двойной связи и т. д. В нашем примере атом **1** (атом углерода) связан с углеродным атомом **2** через одинарную связь и с атомами водорода **4, 5 и 6** также через одинарную связь.

Кроме того, продолжающаяся таблица содержит избыточную информацию, которую можно исключить для простоты записи. Помимо дублирующей записи, атомы водорода можно также опустить в «обычных» органических соединениях, при этом полезная информация не теряется (рис. 15).



Атомный индекс	Элемент	1-й индекс атома	Порядок связи	2-й индекс атома	Порядок связи
1	C	2	1		
2	C			3	2
3	O				



Атомный индекс	Элемент	1-й индекс атома	Порядок связи
1	C	2	1
2	C	3	2
3	O		

Рис. 15. Неизбыточная продолжающаяся таблица этанола (рассматриваются только неводородные атомы, связи с самыми низкими индексами вычисляются один раз)

Почти все химические информационные системы работают со своим собственным типом продолжающихся таблиц. Они часто используют разные форматы для внутренних и внешних продолжающихся таблиц. Во многих случаях внутренние продолжающиеся таблицы являются дублирующими, тем самым значительно увеличивается скорость обработки данных. Внешние продолжающиеся таблицы обычно являются избыточными для сохранения пространства информации на диске (рис.16).

Первая линия продолжающейся таблицы определяет, из какого количества атомов построена данная молекула (9) и какое количество связей имеется в данном соединении (8).

Все атомы описываются подробно в **атомном блоке** (строки 2–10). Каждый атом представлен в виде одного ряда, который определяет декартовы координаты и атомный символ.

После атомного блока идет **блок связей** (строки 11–18). Каждая линия этого блока определяет, какие два атома связаны между собой, мультиплетность связи и стереоконфигурацию связи. Первая колонка указывает первый атом, вторая колонка — атом, с которым

Количество атомов	Количество связей	Координата X	Координата Y	Координата Z	
9	8	-0.1602	-1.3065	0.0000	C
		-0.1602	0.2334	0.0000	C
		1.1879	0.7100	0.0000	O
		-1.1879	-1.6699	0.0000	H
		0.3536	-1.6699	-0.8900	H
		0.3536	-1.6699	0.8900	H
		-0.6741	0.5967	-0.8900	H
		-0.6741	0.5967	0.8900	H
		1.1879	1.6699	0.0000	H
1	2	1	1		
1	4	1	1		
1	5	1	4		
1	6	1	7		
2	3	1	1		
2	7	1	4		
2	8	1	7		
3	9	1	1		

} атомы

} связи

Рис. 16. Представление строения молекулы в виде продолжающейся таблицы

связан первый атом, в третьей колонке указан тип связи (одинарная — 1, двойная — 2, тройная — 3), в четвертой — стереохимические особенности связи.

Таким образом, проанализировав продолжающуюся таблицу, можно нарисовать молекулу. В приведенном примере зашифрована молекула этанола.

В табл. 13 приводятся наиболее общие форматы файлов для информации о химической структуре и их возможности представления или кодирования строения, конфигурации, т. е. стереохимии, и 3D-структуры или конформации. Большинство из этих форматов используют декартовы координаты для представления соединения в трехмерном пространстве.

Общие форматы файлов для представления химических структур

Формат файла	Представление (кодирование)		
	строение	конфигурация	3D-структура / конформация
MDL SDfile	Да	Да	Да
SMILES	Да	Да	Нет
SYBYL MOL2	Да	Неясно	Да
PDB	Да	Неясно	Да
XYZ	Нет	Нет	Да
Z-МАТРИЦА	Нет	Неясно	Да

3D-структура молекулы может быть получена из экспериментальных данных или при помощи вычислительных методов. Существуют программы, которые способны предсказывать без вмешательства пользователя трехмерную молекулярную модель, исходя из строения и стереохимической информации молекулы. Такие программы называются автоматическими генераторами 3D-структуры (рис. 17).

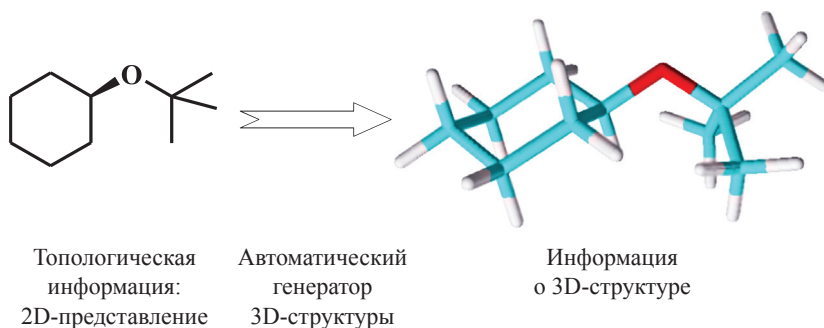


Рис. 17. Автоматическая генерация 3D-структуры

Кодирование структуры

Второй тип внешнего представления структур химических соединений и реакций между ними основан на непосредственном кодировании матрицы смежности молекулярного графа. Такие распространенные форматы, как MOL, SDF и RDF, которые в настоящее время являются стандартными для обмена химической информацией, можно считать способами представления в виде текстового файла матрицы смежности молекулярного графа (табл. 14).

Таблица 14

Наиболее важные форматы файлов для обмена информацией о химической структуре

Формат файла	Суффикс	Комментарии	Поддержка
MDL Mol-file	*.mol	Мол-файл, наиболее широко используемый табличный формат	www.mdli.com
SD-file	*.sdf	Файл с данными о структуре (расширение MDL мол-файла), содержащий одно или несколько соединений	www.mdli.com
RD-file	*.rdf	Файл с данными о реакции, (расширение MDL мол-файла), содержащий одну реакцию или набор из нескольких реакций	www.mdli.com
SMILES	*.smi	SMILES, наиболее широко используемый линейный код и формат файла	www.daylight.com
PDB-file	*.pdb	Файл банка данных о белках, содержащий информацию о 3D-структуре белков и полинуклеотидов	www.rcsb.org
CIF	*.cif	Файл с кристаллографической информацией и данными о 3D-структуре органических молекул	www.iucr.org/iucr-top/cif

Формат файла	Суффикс	Комментарии	Поддержка
JCAMP	*.jdx *.dx *.cs	Совместный файл с атомными и молекулярными физическими данными, а также со структурными и спектроскопическими данными	www.jcamp.org
CML	*.cml	Язык химической разметки, расширение XML со специализацией в химии	www.xml-cml.org

Этой же цели служат и специфические форматы MOL2, HIN, РСМ и др., предназначенные для работы с распространенными программами по молекулярному программированию.

Существует множество типов форматов файлов для хранения информации о химических структурах. Тем не менее только некоторые из них широко принимаются в хемоинформационном обществе и используются в качестве стандартных форматов для обмена информацией о химических структурах и реакциях.

Наиболее широко распространенными являются два формата, а именно Mol-file и SD-file, которые были впервые описаны группой Dalby из компании MDL (Molecular Design Limited).

3.3. МОЛ-ФАЙЛ (MOL-FILE)

Мол-файл описывает одну молекулярную структуру, которая также может состоять из фрагментов молекулы. Рассмотрим мол-файл на примере молекулы ацетальдегида (рис. 18).

Каждый мол-файл состоит из двух частей: строки параметров — так называемой «шапки», строго специфичной для мол-файлов (строки 1–3), и продолжающейся таблицы (строки 4–18), которая составляет основу для такого вида форматов.

Первая строка «шапки» содержит информацию о названии молекулы. Если название по каким-либо причинам отсутствует, то эта строка остается незаполненной. Например, для ацетальдегида

указаны два названия: его идентификационный номер из NCI базы данных (NSC 7594) и название соединения по ИЮПАК.

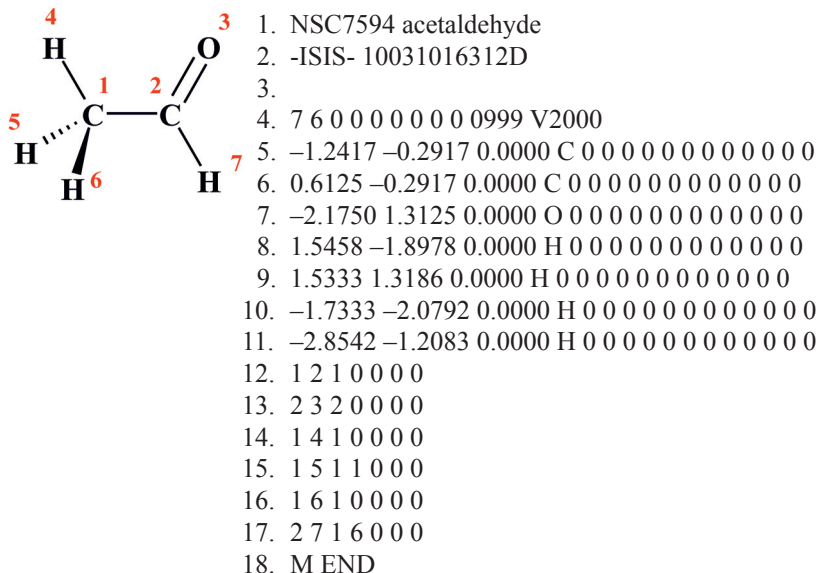


Рис. 18. Мол-файл молекулы ацетальдегида

Вторая строка имеет более строгий формат и содержит общую информацию об имени пользователя, о программе, которая используется для получения данного файла, и о дате и времени, когда этот файл был создан. Информация о дате и времени формируется из ряда двузначных значений, представляющих месяц (**10** в приведенном примере), дату (**03**), год (**10**), час (**16**) и минуты (**31**) соответственно. Также указываются атомные координаты (**2D** или **3D**).

Третья строка «шапки» обычно остается пустой или содержит комментарии.

Строки 4–18 образуют продолжающуюся таблицу, содержащую описание атомов, составляющих данное соединение, которые могут быть полностью или частично соединены связями. Такой набор атомов может представлять молекулу, фрагмент молекулы, субструктуры и т. д. В случае мол-файла блок продолжающейся таблицы описывает одну молекулу.

Первая линия продолжающейся таблицы называется **линией подсчета (расчета)**, она определяет, из какого количества атомов построена данная молекула (7), какое количество связей имеется в данном соединении (6), является ли молекула хиральной (1 — хиральна, 0 — нет) и т. д. Далее может идти перечисление каких-либо других свойств. Последняя позиция в этой строке указывает версию формата продолжающейся таблицы, используемого в данном файле. В приведенном примере это версия **V2000**. Также существует более новая, расширенная версия V3000.

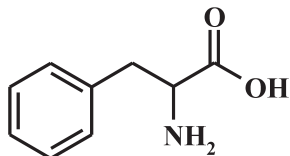
Все семь атомов, которые указаны в строке подсчета, описываются подробно в **атомном блоке** (строки 5–11). Каждый атом представлен в виде одного ряда, который определяет декартовы координаты, атомный символ, отклонение от атомной массы основного изотопа по сравнению с периодической таблицей, заряд, стереохимические особенности атома, число отдельно изображенных атомов водорода, нестандартную валентность и ряд других свойств. Декартовы координаты определяют модель молекулы (2D или 3D), как указано во второй строке файла. 2D-модели могут быть получены, например, при помощи программы ISIS/DRAW. 3D-структурные данные получаются в результате экспериментов или теоретических вычислений (например, программа CORINA). 3D-координаты можно увидеть в третьей колонке атомного блока (Z-координаты). Если эта колонка содержит только значения нуля, то мол-файл содержит только 2D-координаты.

После атомного блока идет **блок связей** (строки 12–17). Каждая линия этого блока определяет, какие два атома связаны между собой, мультиплетность связи и стереоконфигурацию связи. Первая колонка указывает первый атом, вторая колонка — атом, с которым связан первый атом, в третьей колонке указан тип связи (одинарная — 1, двойная — 2, тройная — 3), в четвертой — стереохимические особенности связи.

Одинарная связь содержит значение 0, если нет стереоспецифичности (обычная связь, в плоскости), значение 1 — над плоскостью, 2 — за плоскостью, 4 — вверх, 5 — направление неизвестно, 6 — вниз. *Цис/транс* или *E/Z* конфигурация двойной связи определяется *x*, *y*, *z* координатами атомного блока, если

значение **0**. Если указано значение **3**, значит, это двойная связь — либо *цис*, либо *транс*.

В качестве более сложного примера можно привести мол-файл фенилаланина (рис. 19).



-ISIS- 10030708132D

Комментарии

12 12 0 0 0 0 0 0 0 0999 V2000

```

-0.7764 -2.3791 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.7776 -3.2065 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.0628 -3.6194 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.6537 -3.2060 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.6508 -2.3755 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.0646 -1.9664 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1.3637 -1.9603 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.0797 -2.3701 0.0000 C 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 2.7927 -1.9549 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.0829 -3.1951 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3.5087 -2.3647 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.7895 -1.1299 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```

```

6 1 1 0 0 0 0
1 2 2 0 0 0 0
5 7 1 0 0 0 0
3 4 2 0 0 0 0
7 8 1 0 0 0 0
8 9 1 0 0 0 0
4 5 1 0 0 0 0
8 10 1 1 0 0 0
2 3 1 0 0 0 0
9 11 1 0 0 0 0
5 6 2 0 0 0 0
9 12 2 0 0 0 0

```

M END

Рис. 19. Мол-файл молекулы фенилаланина

Последняя часть файла, представленного здесь, — это **блок**, который может содержать разнообразные свойства. Однако в большинстве случаев этот блок остается пустым, за исключением последней строки (строка 18).

3.4. PDB-ФАЙЛЫ (БАЗА ДАННЫХ ПО БЕЛКАМ)

PDB — это база для хранения данных по трехмерной структуре больших биологических молекул, таких как белки и нуклеиновые кислоты (табл. 15). Данные получают при помощи рентгеноструктурного анализа или спектроскопии ядерного магнитного резонанса (ЯМР); их достоверность подтверждают биологи и биохимики по всему миру.

PDB является ключевым источником для структурной биологии (структурная геномика).

Таблица 15

База данных по белкам (PDB)

Экспериментальный метод	Белки	Нуклеиновые кислоты	Комплексы «белок — нуклеиновая кислота»	Другое	Сумма
Дифракция рентгеновских лучей	52464	1205	2431	17	56117
ЯМР-спектроскопия	7300	898	155	7	8360
Электронная микроскопия	196	17	76	0	289
Гибриды	18	1	1	1	21
Другое	124	4	4	13	145
<i>Итого</i>	60102	2125	2667	38	64932

Формат базы данных по белкам обычно используется для белков, но он может также использоваться для других типов молекул. Формат фиксированной ширины колонки имеет встроенное

максимальное количество атомов, остатков и цепей; это в настоящее время приводит к расщеплению очень больших структур, таких как рибосомы, на мультиплетные файлы (например, 3IIM, 3IIN, 3IIO, 3IIP). Некоторые файлы PDB содержат дополнительный раздел, описывающий возможность соединения атома и его положение. Поскольку эти файлы иногда используются, чтобы описать макромолекулярные скопления или молекулы, представленные в определенном растворителе, они могут быть очень большими и часто приходится их сжимать. Некоторые инструментарии, такие как Jmol и KiNG, могут прочитать файлы PDB в формате *gzipped*. PDB поддерживает технические требования формата файла PDB и его альтернативы XML, PDBML. В августе 2007 г. было произведено довольно существенное изменение в спецификации формата PDB (к версии 3.0) и исправление многих проблем файла в существующей базе данных. Типичное расширение для файла PDB — *.pdb*, хотя некоторые более старые файлы используют *.ent* или *.brk*.

Эти данные показывают, что строение большинства структур определяется рентгеноструктурным анализом, но в то же самое время приблизительно 15 % структур в настоящее время определяются спектроскопией ядерного магнитного резонанса белка, а некоторые — даже крио-электронной микроскопией.

Формат файла

Формат файла, первоначально используемый PDB, назвали форматом файла PDB. Этот оригинальный формат был ограничен шириной компьютерных перфокарт до 80 характеристик на линию. Приблизительно в 1996 г. начал постепенно вводиться макромолекулярный кристаллографический информационный формат файла — mmCIF. В 2005 г. была описана версия XML этого формата, названная PDBML. Структурные файлы могут быть загружены в любой из этих трех форматов. Фактически отдельные файлы легко загружаются в графические пакеты, используя веб-адреса:

— например, для файлов формата PDB — [http:// www.pdb.org/pdb/files/4hhb.pdb.gz](http://www.pdb.org/pdb/files/4hhb.pdb.gz);

— для PDBML(XML)-файлов — [http:// www.pdb.org/pdb/files/4hhb.xml.gz](http://www.pdb.org/pdb/files/4hhb.xml.gz).

«4hbb» — идентификатор PDB. Каждая структура, изданная в PDB, получает четырехсимвольный алфавитно-цифровой идентификатор, его ID номер PDB. (Он не может использоваться в качестве идентификатора для биомолекул, потому что часто несколько структур для той же самой молекулы — в различной окружающей среде или конформациях — содержатся в PDB с различными ID PDB.)

3.5. Z-МАТРИЦА

Следующим методом для представления молекулы в 3D-пространстве является использование **внутренних координат**, таких как длины связей, углы связей и двугранные углы. Двугранный угол образуется из четырех атомов и помогает определить размер молекулы. Под связью в данном случае подразумевают не химическую связь, а просто вектор, направленный от одного атома к другому, хотя они могут и совпадать.

Наиболее широко используемый способ описания молекулы с помощью ее внутренних координат — это так называемая **Z-матрица**. Z-матрица называется так потому, что второй атом всегда располагается вдоль оси аппликат (оси Z). Внутренние координаты описывают пространственное расположение атомов относительно других атомов (рис. 20).

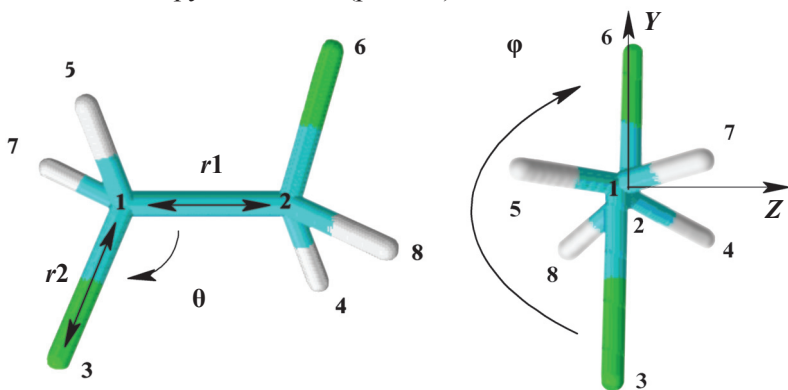


Рис. 20. Внутренние координаты 1,2-дихлорэтана:
 r_1 и r_2 — длины связей; θ — угол связи; ϕ — двугранный угол

Сущность этого способа представления координат состоит в конструировании для каждого задаваемого атома своей собственной полярной системы координат. Для этой цели используются любые три атома, положение которых было определено ранее. Эти атомы называются *базисными*, причем первый базисный атом принимается за центр полярной системы координат, второй базисный атом задает базовую ось, а третий базисный атом — базовую плоскость полярной системы координат (рис. 21).

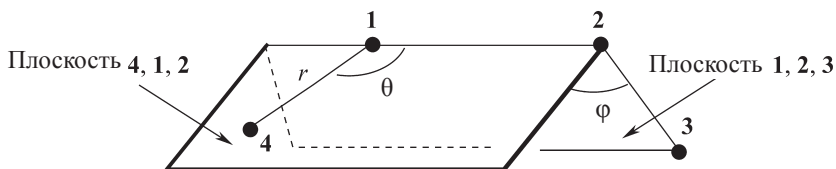


Рис. 21. Молекулярные координаты:

r — длина связи; θ — валентный угол связи; φ — двугранный угол

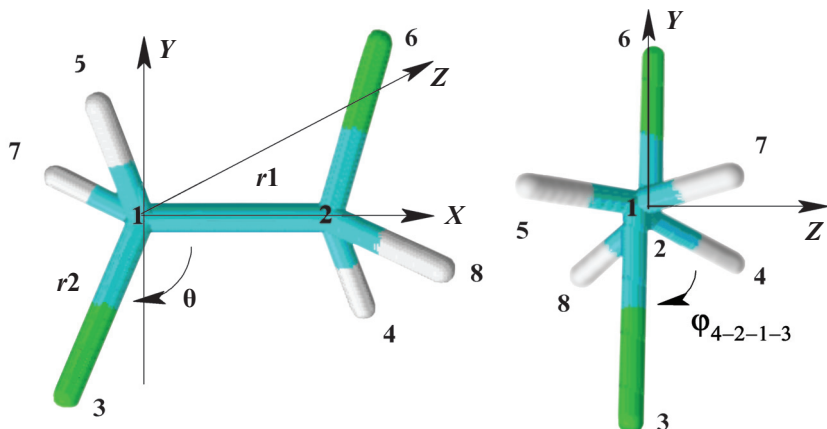
Для однозначного определения положения атома 4 в пространстве необходимо определить совокупность трех координат: радиус-вектора r , плоского угла θ и двугранного угла φ . Как видно из схемы, величина r находится из расстояния между задаваемым атомом 4 и базисным атомом 1, плоский угол θ — суть угол между атомами 4, 1 и 2 с вершиной на базисном атоме 1, двугранный (межплоскостной, диэдральный, торсионный) угол φ — это угол между плоскостями 4, 1, 2 и 1, 2, 3, т. е. угол между перпендикулярами к базовой оси, опущенными из одной точки в обе плоскости.

Из сказанного следует, что нельзя использовать базисные атомы, лежащие на одной прямой, так как в этом случае базовая плоскость 1, 2, 3 однозначно не определена и, следовательно, угол φ не может быть рассчитан. По этой же причине следует выбирать базисные атомы 1 и 2 таким образом, чтобы угол θ (4–1–2) не был равен 0 или 180° , так как в этом случае однозначно не определяется плоскость 4, 1, 2.

Задачей пользователя является разумный выбор базисных атомов с целью приближенного определения величин r , θ и φ . В ходе

последующего расчета программа найдет оптимальные значения геометрических параметров.

На рис. 22 показана Z-матрица 1,2-дихлорэтана.



C1						
C2	1.5	1				
C13	1.7	1	109	2		
H4	1.1	2	109	1	-60	3
H5	1.1	1	109	2	180	4
C16	1.7	2	109	1	60	5
H7	1.1	1	109	2	-60	6
H8	1.1	2	109	1	180	7

Рис. 22. Z-матрица 1,2-дихлорэтана

Каждая линия в Z-матрице представляет один атом молекулы. В первой линии атом **1** определяется как **C1**, т. е. этот углеродный атом лежит в начале координатной системы. Второй атом **C2** находится на расстоянии 1,5 Å (вторая колонка) от атома **1** (третья колонка) и должен всегда помещаться на главной оси (X-ось на рис. 22). Третий атом (хлор) **C13** должен лежать на XY-плоскости;

он находится на расстоянии 1,7 Å от атома 1, и угол θ между атомами 3–1–2 составляет 109° (четвертая и пятая колонка). Третий тип внутренней координаты (двугранный угол ϕ) вводится на четвертой линии Z-матрицы в шестой и седьмой колонках. Это угол между плоскостями, которые образованы атомами 4, 2, 1 и 2, 1, 3 (XY-плоскость). За исключением первых трех атомов, каждый атом описывается набором трех внутренних координат: 1) расстояние от предыдущего определенного атома; 2) угол связи, образованной атомом с двумя предыдущими атомами; 3) двугранный угол атома, образованный с тремя предыдущими атомами. Число ($3N - 6$) внутренних координат, где N — число атомов в молекуле, представляет структуру в 3D-пространстве, а также соответствует числу степеней свободы молекулы.

Z-матрица может быть преобразована в декартовы координаты. Также возможно и обратное преобразование без потери информации, как показано ниже на примере метана.

Молекула метана в декартовых координатах (в ангстремах, Å):

```

C 0.000000 0.000000 0.000000
H 0.000000 0.000000 1.089000
H 1.026719 0.000000 -0.363000
H -0.513360 -0.889165 -0.363000
H -0.513360 0.889165 -0.363000

```

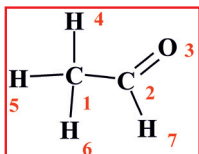
Молекула метана в виде Z-матрицы:

```

C
H 1 1.089000
H 1 1.089000 2 109.4710
H 1 1.089000 2 109.4710 3 120.0000
H 1 1.089000 2 109.4710 3 -120.0000

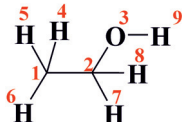
```

Разберем представление разных молекул в виде Z-матриц. Внешний вид этих матриц также может отличаться, однако сущность этого метода не изменяется. В табл. 16 приводится изображение Z-матрицы этанала.



№ п/п	Угол	λ , Å	Номер, от которого измеряется длина связи (λ)	Валентный угол	Номер, от которого измеряется валентный угол	Двугранный угол	Номер, от которого измеряется двугранный угол
1	C	0		0		0	
2	C	1.49	1	0			
3	O	1.22	2	120	1		
4	H	3.00	1	116	2	5	3
5	H	1.09	1	116	2	60	3
6	H	1.09	1	116	2	-60	3
7	H	1.09	2	120	3	0	1

Другой пример Z-матрицы этанола приведен на рис. 23. Единицы и нули между координатами обозначают, необходимо или нет оптимизировать данную координату: 1 — оптимизировать, 0 — не оптимизировать.



Name:	MOPAC file created on 9/2 15:15:47 1994 by HYPERCHEM									
C	00000.0000	0	00000.0000	0	00000.0000	0	0	0	0	0
C	00001.5399	1	00000.0000	0	00000.0000	0	1	0	0	
O	00001.4299	1	00109.4710	1	00000.0000	0	2	1	0	
H	00001.0900	1	00109.4710	1	00179.9999	1	1	2	3	
H	00001.0900	1	00109.4710	1	00299.9994	1	1	2	3	
H	00001.0900	1	00109.4714	1	00059.9997	1	1	2	3	
H	00001.0900	1	00109.4709	1	00299.9994	1	2	1	4	
H	00001.0900	1	00109.4714	1	00059.9997	1	2	1	4	
H	00000.9599	1	00109.4710	1	00060.0005	1	3	2	7	
0										

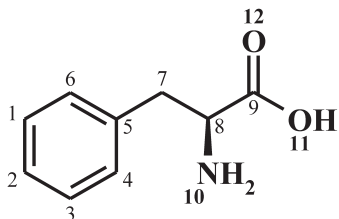
↑
↑
↑
↑
↑
↑
↑
↑

Координата 1
Координата 2
Координата 3
Связь с атомом
Валентный угол

Двугранный угол

Рис. 23. Z-матрица молекулы этанола

В качестве другого примера можно привести более сложную молекулу фенилаланина (рис. 24).



Name :

MOPAC file created on 3/10 13:19:25 2007 by HYPERCHEM

C	0.0000	0	000.0000	0	000.0000	0	0	0	0
C	1.4000	1	000.0000	0	000.0000	0	1	0	0
C	1.4000	1	119.9999	1	000.0000	0	2	1	0
C	1.4000	1	119.9999	1	000.0000	1	3	2	1
C	1.3999	1	119.9999	1	000.0000	1	4	3	2
C	1.3999	1	120.0000	1	000.0000	1	5	4	3
C	1.5200	1	119.9999	1	179.9999	1	5	4	3
C	1.5399	1	109.4709	1	000.0000	1	7	5	6
C	1.5200	1	109.4710	1	179.9999	1	8	7	5
N	1.4700	1	109.4710	1	299.9994	1	8	7	5
O	1.3599	1	120.0000	1	180.0000	1	9	8	7
O	1.2199	1	119.9999	1	240.0005	1	9	8	10
H	1.0799	1	120.0000	1	179.9999	1	1	2	3
H	1.0799	1	120.0000	1	179.9999	1	2	1	6
H	1.0799	1	120.0000	1	179.9999	1	3	2	1
H	1.0800	1	119.9999	1	179.9999	1	4	3	2
H	1.0799	1	120.0000	1	179.9999	1	6	5	4
H	1.0900	1	109.4710	1	119.9994	1	7	5	6
H	1.0899	1	109.4714	1	239.9997	1	7	5	6
H	1.0900	1	109.4714	1	300.0002	1	8	7	18
H	1.0100	1	109.4710	1	299.9994	1	10	8	9
H	1.0100	1	109.4709	1	179.9991	1	10	8	20
H	0.9599	1	109.4709	1	000.0000	1	11	9	12
O									

Рис. 24. Z-матрица молекулы фенилаланина

3.6. ТЕОРИЯ ГРАФОВ (ДИАГРАММЫ ГРАФОВ)

Аналогия между структурной диаграммой и топологическими диаграммами является основой для разработки теоретического алгоритма графов по представлению информации о химической структуре.

Понятие «молекулярный граф» является базовым для компьютерной химии и хемоинформатики. Как и структурная формула, молекулярный граф является моделью молекулы, и как всякая модель, он отражает далеко не все свойства прототипа. В отличие от структурной формулы, где всегда указывается, к какому химическому элементу относится данный атом, вершины молекулярного графа могут быть непомяченными — в этом случае молекулярный граф будет отражать только структуру, но не состав молекулы. Точно так же ребра молекулярного графа могут быть непомяченными — в таком случае не будет делаться различие между ординарными и кратными химическими связями. В некоторых случаях может использоваться молекулярный граф, отражающий только углеродный скелет молекулы органического соединения. Такой уровень абстрагирования удобен для вычислительного решения широкого круга химических задач.

В математических терминах структурные диаграммы, которые рисуются химиками, можно рассматривать как обычные графы. Графы состоят из узловых точек (вершин), которыми являются атомы, и ребер, которые представляют собой связи. В органической химии эти графы чаще всего упрощаются путем представления углеродных атомов в качестве точек, где соединяются линии, и ребер (связей) (рис. 25).

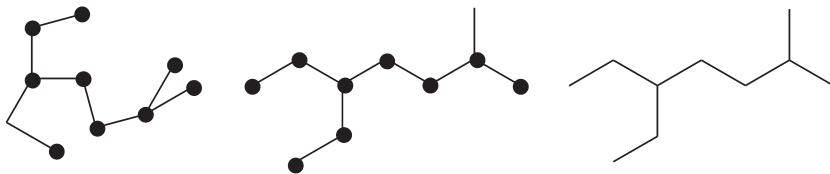


Рис. 25. Различные представления идентичной диаграммы в теории графов

В теории графов важны только связи, а длины ребер или углы между ними не имеют никакого значения.

Такая диаграмма называется топологической диаграммой, так как она показывает только связи между атомами и тип связей между ними. Этот тип структурной презентации не содержит данных о 3D-структуре, топографии молекулы.

Обычно структурная диаграмма является ненаправленной (связи не имеют направления) и маркированной (узловые точки характеризуются символами атомов). В теории графов диаграмма не несет информации о геометрии молекулы. Диаграмма плотности (весовая) содержит номера или символы, которые ставятся в соответствие узловым точкам. Две узловые точки могут иметь различные ребра (в химии, например, это многочисленные связи) (рис. 26).

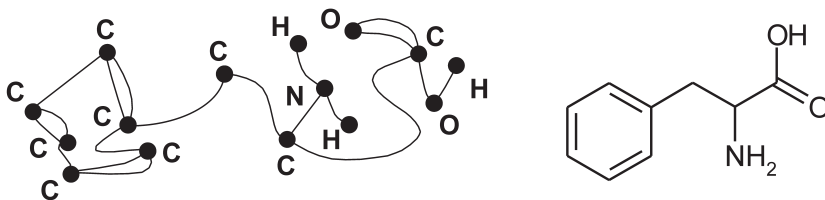


Рис. 26. Представление фенилаланина в теории графов (маркированная диаграмма плотности с различными типами атомов и связей)

Естественным расширением молекулярного графа является реакционный граф, ребра которого соответствуют образованию, разрыву и изменению порядка связей между атомами.

Графы также можно представить как **матрицы**. Их главное преимущество состоит в том, что вычисление цепей и циклов можно легко выполнить при помощи хорошо известных операций с матрицами.

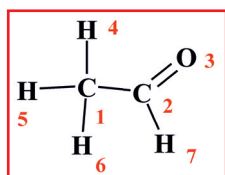
Матрица структуры с n атомами состоит из массива (таблицы), включающего $n \times n$ записей. Молекула с различными типами атомов и связей может быть представлена в матричной форме разными способами в зависимости от вида записи, который выбирается для атомов и связей. Таким образом, было предложено огромное количество матриц.

В тех матрицах, которые будут рассмотрены ниже, атомы водорода не всегда указаны, так как их номера и позиции могут быть вычислены из органической структуры на основании правил валентности других атомов.

Другой особенностью в описании матриц является то, что каждый атом описывается дважды – в колонке и в ряду. Матрицы, в которых все элементы описываются дважды, называются избыточными (дублирующими). В неизбыточной матрице каждый элемент записывается только один раз.

3.6.1. Матрица смежности

Матрица смежности молекулы, состоящей из n атомов, представляет собой квадратную матрицу ($n \times n$) с записями, которые показывают все взаимодействия атомов. Пересечение ряда и колонки дает значение **1**, если соответствующие атомы связаны. Если между рассматриваемыми атомами никаких связей не наблюдается, то положение в матрице дает значение **0**. Это представление матрицы называется матрицей Boolean с двоичными знаками (**0** или **1**) (рис. 27).

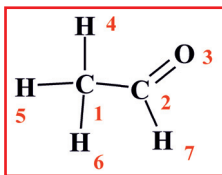


$n = 7$

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

Рис. 27. Матрица смежности (7×7) этаноля

Как видно из рис. 27, диагональные элементы матрицы всегда равны нулю. Таким образом, эта матрица является избыточной и может быть уменьшена наполовину (рис. 28). Для пояснения все записи нулей пропускаются.



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

a

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

б

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

в

	1	2	3
1		1	
2			1
3			

г

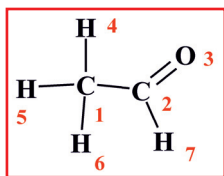
Рис. 28. Избыточная матрица смежности этанала (*a*), упрощенная при помощи: *б* — удаления нулевых значений; *в* — устранения дублей; *г* — удаления информации об атомах водорода

При таком представлении матрицы пространство памяти зависит только от числа узловых точек (атомов) и не зависит от числа связей. Как видно из рис. 28, вся необходимая информация матрицы

смежности может быть найдена в избыточной матрице. Но матрица смежности не подходит для реконструкции структуры молекулы, так как не предоставляет информации о порядке связей.

3.6.2. Матрица расстояний

Элементы данного вида матрицы содержат значения, которые определяют самое короткое расстояние между связанными атомами. Расстояния можно изображать или как геометрические расстояния (в ангстремах), или как топологические (подсчет числа связей по кратчайшему пути) (рис. 29).



	C1	C2	O3	H4	H5	H6	H7
C1	0	1.400	2.190	1.022	1.023	1.022	2.106
C2	1.400	0	1.123	1.999	1.982	1.999	1.022
O3	2.190	1.123	0	2.349	2.708	2.995	1.859
H4	1.022	1.999	2.349	0	1.668	1.661	2.895
H5	1.023	1.982	2.708	1.668	0	1.668	2.562
H6	1.022	1.999	2.955	1.661	1.668	0	2.336
H7	2.106	1.022	1.859	2.895	2.566	2.336	0

a

	C1	C2	O3	H4	H5	H6	H7
C1	0	1	2	1	1	1	2
C2	1	0	1	2	2	2	1
O3	2	1	0	3	3	3	2
H4	1	2	3	0	2	2	3
H5	1	2	3	2	0	2	3
H6	1	2	3	2	2	0	3
H7	2	1	2	3	3	3	0

b

Рис. 29. Матрицы расстояния этанола:

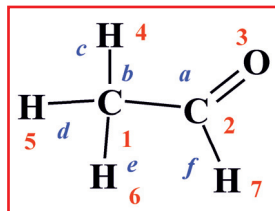
a — геометрические расстояния, Å;

b — топологические расстояния

3.6.3. Матрица понижения

Матрица понижения представляет собой $n \times m$ матрицу, где узловые точки (атомы) определяют колонки (n), а ребра (связи) соответствуют рядам (m). Запись предоставляет значение **1**, если соответствующее ребро заканчивается в этой определенной узловой точке (рис. 30). В неквадратной матрице атомы указаны в колонках, связи – в рядах.

	C1	C2	O3	H4	H5	H6	H7
<i>a</i>	1	1	0	0	0	0	0
<i>b</i>	0	1	1	0	0	0	0
<i>c</i>	1	0	0	1	0	0	0
<i>d</i>	1	0	0	0	1	0	0
<i>e</i>	1	0	0	0	0	1	0
<i>f</i>	0	1	0	0	0	0	1



$$n = 7, m = 6$$

a

	C1	C2	O3	H4	H5	H6	H7
<i>a</i>	1	1					
<i>b</i>		1	1				
<i>c</i>	1			1			
<i>d</i>	1				1		
<i>e</i>	1					1	
<i>f</i>		1					1

	C1	C2	O3
<i>a</i>	1	1	
<i>b</i>		1	1

b

b

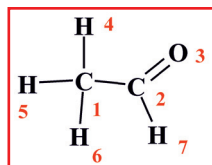
Рис. 30. Матрицы понижения молекулы этаноля:

a — избыточная матрица понижения этаноля; уменьшенная матрица:
b — путем удаления значений нуля; *b* — путем удаления информации об атомах водорода

3.6.4. Матрица связей

Матрица связей по виду похожа на матрицу смежности, но в отличие от нее дает информацию еще и о порядке связей атомов. Элементы матрицы получают значение **2**, если существует двойная связь между атомами (например, атомы **2** и **3** в указанном примере). Остальные элементы могут иметь значения **0**, **1** или **3** для других комбинаций связей. Это представление является избыточным (рис. 31).

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



a

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

б

	1	2	3
1		1	
2			2
3			

в

Рис. 31. Матрицы связей молекулы этанала:

a — избыточная матрица связей этанала, в которой удалены значения нуля;
б — матрица, сжатая путем устранения дублей; *в* — простое представление неизбыточной матрицы за счет удаления атомов водорода

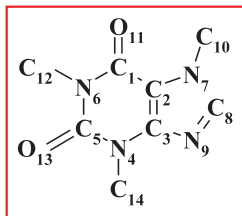
Матрица связей дает такое представление молекулярной структуры, в котором каждый атом, каждая связь и тип связи кодируются по отдельности и всегда обозначаются явным образом. В отличие от линейных номенклатур, правила составления матриц связей просты и легко применимы для кодирования структур любой сложности.

Главная диагональ матрицы связей включает коды атомов структуры. Каждому типу атомов соответствует свой код. Недиагональный элемент матрицы связей a_{ij} содержит информацию о связи между i -м и j -м атомами и является кодом данного типа связи. Правила кодирования структур в виде матриц связей легко понять с помощью следующего примера.

На рис. 32 приведена структурная формула кофеина и соответствующая ей матрица связей. Также приведены числовые коды различных типов атомов и химических связей. Последовательность нумерации атомов в структуре произвольна и соответствует последовательности расположения кодов атомов на главной диагонали матрицы связей. Атомы водорода не включены, так как их расположение легко рассчитать с помощью правил валентности. Равенство недиагонального элемента a_{ij} нулю означает отсутствие связи между i -м и j -м атомами.

Использованные в рассмотренном примере числовые коды выбраны произвольно и не являются общепринятым стандартом. Впрочем, это обстоятельство не создает дополнительных трудностей, так как любая матрица связей легко может быть переведена в другую кодировку с помощью несложной вычислительной процедуры. Кстати, системы линейной кодировки таким свойством не обладают.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1				1					2			
2	1	1	2				1							
3		2	1	1					1					
4			1	3	1									1
5				1	1	1							2	
6	1				1	3						1		
7		1					3	1		1				
8							1	1	2					
9			1					2	3					
10							1			1				
11	2										2			
12						1						1		
13					2								2	
14				1										1



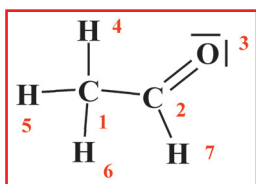
Коды для матрицы связей

Тип атома	Числовые коды	Тип связи	Числовые коды
C	1	Простая	1
O	2	Двойная	2
N	3	Тройная	3
S	4	Ароматическая	4
F	5	Делокализованная	5
Cl	6	Ионная	6
Br	7		
I	8		
P	9		

Рис. 32. Пример матрицы связей для молекулы кофеина

3.6.5. Матрица «связь — электрон» (ВЕ-матрица)

Матрица «связь — электрон» (ВЕ-матрица) была предложена в модели Dugundji-Ugi. Ее можно рассматривать как расширение матрицы связей. Этот тип матрицы дает дополнительное число свободных валентных электронов на соответствующем атоме в диагональных элементах (например, $O_3 = 4$, см. рис. 33).



$$n = 7$$

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

a

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

б

	1	2	3
1		1	
2			2
3			4

в

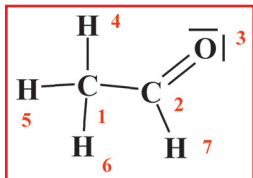
Рис. 33. ВЕ-матрица молекулы этанала:

a — избыточная ВЕ-матрица этанала с пропущенными значениями нуля;
б — матрица, сжатая путем удаления дублей; *в* — простая презентация избыточной матрицы за счет удаления атомов водорода

В сущности, ВЕ-матрица иллюстрирует все валентные электроны атомов в молекуле, но одни из них включены в связи, а другие

считаются свободными электронами с атомом. ВЕ-матрица имеет ряд интересных математических свойств, которые напрямую отображают химическую информацию.

В примере на рис. 34 атом кислорода имеет: $2 + 4$ (ряд) $+ 2 + 4$ (колонка) $- 4$ (диагональный элемент) $= 8$ электронов. Это показывает, что атом кислорода подчиняется октетному правилу.



$$n = 7$$

	1	2	3	4	5	6	7	Сумма в ряду	Элемент
1		1		1	1	1		4	С
2	1		2				1	4	С
3		2	4					6	О
4	1							1	Н
5	1							1	Н
6	1							1	Н
7		1						1	Н
Сумма в колонке	4	4	6	1	1	1	1	36	
Сумма пересечения	8	8	8	2	2	2	2		

Рис. 34. ВЕ-матрица для определения числа валентных электронов (сумма каждого ряда) на атомах (подтверждение октетного правила)

ВЕ-матрица также обеспечивает основу для представления матрицы химических реакций.

3.7. ПРЕДСТАВЛЕНИЕ МАТРИЦЫ ХИМИЧЕСКИХ РЕАКЦИЙ

Такие распространенные форматы, как SDF и RDF, которые в настоящее время являются стандартными для обмена химической информацией, можно считать способами представления в виде текстового файла матрицы смежности молекулярного графа.

3.7.1. SD-файлы

SD-файлы содержат структурную информацию и группы данных для одного и более соединений (рис. 35). Преимуществом этих файлов является обмен данными не только между базами данных, но и между вычислительным программным обеспечением. Большинство таких программ могут писать результаты своих вычислений в этом формате. В SD-файле каждая молекула представляется своим собственным мол-файлом с дополнительными группами данных, описывающими ее неструктурные свойства (молекулярный вес, теплота образования, молекулярные дескрипторы, биологическая активность и т. д.).

Информация по каждой молекуле ограничивается при помощи разделительной линии, содержащей только знаки «\$\$\$\$». Каждая группа данных начинается со строки параметров, где указывается название молекулы. Далее один или более рядов содержат фактические данные, они ограничиваются при помощи пустой строки.

3.7.2. RXN-файлы (реакционные файлы)

RXN-файлы содержат структурные данные для реагентов и продуктов реакции (рис. 36). Первая строка содержит информацию о реагентах и продукте реакции, вторая строка всегда остается пустой. Третья строка содержит информацию о названии и версии программы, в которой был записан файл, дату и время записи, регистрационный номер реакции. Четвертая линия — это линия для комментариев. Если их нет, то эта строка остается незаполненной.

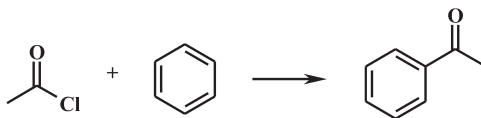
```

-CPSS- 0804941117
13 14 0 0 0 0 0 0 0 0 0 0
 0.8400 -0.1600 0.0000 N 0 0 0 0 0 0 0 0
 1.4800 0.4300 0.0000 N 0 0 0 0 0 0 0 0
 0.0900 0.2700 0.0000 N 0 0 0 0 0 0 0 0
 1.1100 1.2100 0.0000 C 0 0 0 0 0 0 0 0
 0.2700 1.1200 0.0000 C 0 0 0 0 0 0 0 0
 0.8400 -1.0300 0.0000 C 0 0 0 0 0 0 0 0
 1.5300 1.9900 0.0000 C 0 0 0 0 0 0 0 0
 1.0700 2.7400 0.0000 C1 0 0 0 0 0 0 0 0
 1.5900 -1.4600 0.0000 C 0 0 0 0 0 0 0 0
 0.0800 -1.4600 0.0000 C 0 0 0 0 0 0 0 0
 1.5900 -2.3300 0.0000 C 0 0 0 0 0 0 0 0
 0.0700 -2.3200 0.0000 C 0 0 0 0 0 0 0 0
 0.8400 -2.7600 0.0000 C 0 0 0 0 0 0 0 0
 2 1 1 0 2 0 0
 3 1 1 0 2 0 0
 4 2 2 0 2 0 0
 5 3 2 0 2 0 0
 6 1 1 0 2 0 0
 7 4 1 0 2 0 0
 8 7 1 0 2 0 0
 9 6 1 0 1 0 0
10 6 2 0 1 0 0
11 9 2 0 1 0 0
12 10 1 0 1 0 0
13 12 2 0 1 0 0
 4 5 1 0 2 0 0
13 11 1 0 1 0 0
> <Sample Ref.>
OC101-12
> <Melting Point>
41.00 - 43.00
> <B1 Record No.>
304
> <ID>
304
$$$$

```

Рис. 35. Пример SD-файла

Далее записываются серии блоков мол-файлов для каждого реагента и продукта реакции, которые начинаются с разделительной линии. Эти блоки всегда находятся в том же порядке, в котором молекулы записаны в реакции: сначала реагенты, а потом уже продукты реакции.




```

$RXN
      ISIS 060920192042
2 1
$MOL
-ISIS- 06091920422D
  4 3 0 0 0 0 0 0 0 0999 V2000
  -2.4708 -2.1542 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -1.7583 -1.7375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -1.7629 -0.9125 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -1.0416 -2.1460 0.0000 Cl 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  2 3 2 0 0 0 0
  1 2 1 0 0 0 0
  2 4 1 0 0 0 0
M END
$MOL
-ISIS- 06091920422D
  6 6 0 0 0 0 0 0 0 0999 V2000
  3.0902 -1.0666 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3.0891 -1.8940 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3.8039 -2.3069 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  4.5203 -1.8935 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  4.5175 -1.0630 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3.8021 -0.6539 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3 4 2 0 0 0 0
  2 3 1 0 0 0 0
  4 5 1 0 0 0 0
  1 2 2 0 0 0 0
  5 6 2 0 0 0 0
  6 1 1 0 0 0 0
M END
$MOL
-ISIS- 06091920422D
  9 9 0 0 0 0 0 0 0 0999 V2000
  9.0861 -1.0666 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.0849 -1.8940 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.7997 -2.3069 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  10.5162 -1.8935 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  10.5133 -1.0630 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.7979 -0.6539 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  11.2262 -0.6478 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  11.9422 -1.0576 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  11.2231 0.1772 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  1 2 2 0 0 0 0
  5 6 2 0 0 0 0
  6 1 1 0 0 0 0
  3 4 2 0 0 0 0
  5 7 1 0 0 0 0
  2 3 1 0 0 0 0
  7 8 1 0 0 0 0
  4 5 1 0 0 0 0
  7 9 2 0 0 0 0
M END

```

Рис. 36. Пример RXN-файла реакции ацилирования бензола

3.7.3. RD-файлы (файлы «реакция — данные»)

В RD-файлах каждая запись определяет молекулу или реакцию и соответствующие данные по соединениям и условиям реакции (рис. 37). Первая строка идентифицирует файл как RD-файл. Во второй строке идут данные о дате и времени записи файла. Далее записываются блоки по идентификации молекул и реакции.

```

SRDFILE 1
$DATM 10/17/91 10:41
$REMT $RIREG 7439
$RXN

REACCS81 1017911041 7439

2 1
$MOL
REACCS8110179110412D 1 0.00380 0.00000 315
4 3 0 0 0 0 0 0 0 0 0 0

1 4 1 0 0 0 4
$MOL
REACCS8110179110412D 1 0.00371 0.00000 8
6 6 0 0 0 0 0 0 0 0 0 0

5 6 2 0 0 0 2
$MOL
REACCS8110179110412D 1 0.00374 0.00000 255
9 9 0 0 0 0 0 0 0 0 0 0

6 9 2 0 0 0 2

$DTYPE rxn:VARIATION(1) : rxnTEXT (1)
$DATUM CrCl3
$DTYPE rxn:VARIATION(1) : LITTEXT(1)
$DATUH A G Repin, Y Y Hakarov-Zenlyanskii, Zur Russ Fiz-Chim, 44,
p.2360, 1974

$DTYPE rxn:VARIATION(1) : CATALYST(1):REGNO
$DATUH SMFMT $MIREG 688

REACCS81101791104120 1 0.00371 0.00000 0
4 3 0 0 0 0 0 0 0 0 0 0

1 4 1 0 0 0 0
$DTYPE rxn:VARIATION(1) : PRODUCT(1) : YIELD
$DATUM 70.0

$RFMT $RIREG 8410
$RXN

REACCS81 1017911041 8410
2 1 SHOL
...

```

Rxnfile header
 RDfile header
 # Reactant and # Product
 Molfile for first reactant
 First Rxn record
 Molfile for second reactant
 Molfile for product
 Data block for reaction
 Start of next Rxn record

Рис. 37. Пример RD-файла

3.7.4. CML (химический язык разметки)

Следующий тип внешнего представления структур химических соединений основан на технологии XML. Наиболее распространенным языком описания химической информации, опирающимся на эти принципы, является CML.

CML — это новый подход к управлению молекулярной информацией, с использованием таких интернет-инструментов, как XML и Java. Химическая информация традиционно хранится во многих различных типах файла, которые запрещают повторное использование документов. CML использует мобильность XML, чтобы помочь разработчикам CML и химикам создать документы, имеющие возможность взаимодействовать (с другой сетью, устройством). Есть много инструментов, которые могут произвести, обработать и рассмотреть документы CML. Издатели могут распределить химическую информацию в пределах документов XML при использовании CML.

CML способен к поддержке широкого диапазона химических понятий, таких как:

- молекулы;
- реакции;
- спектры и аналитические данные;
- компьютерная химия;
- химическая кристаллография и материалы.

Пример химической записи в формате CML:

```
<?xml version="1.0"?>
<document>
<cml title="arotene" id="cml_carotine_karne"
  xmlns="x-schema:cml_schema_ie_02.xml">
  <molecule title="arotene" id="mol_carotine_karne">
    <atomArray>
      <atom id="arotene_karne_a_1">
        <float builtin="x2" units="A">17.3280</float>
        <float builtin="y2" units="A">2.0032</float>
        <string builtin="elementType">C</string>
      </atom>
      ... many atoms deleted, for space ...
    </atomArray>
```

```
<bondArray>
  <bond id="arotene_karne_b_1">
    <string builtin="atomRef">arotene_karne_a_1</string>
    <string builtin="atomRef">arotene_karne_a_2</string>
    <string builtin="order" convention="MDL">2</string>
  </bond>
  ... many bonds deleted, for space ...
</bondArray>
</molecule>
</cml>
</document>
```

4. НУМЕРАЦИЯ АТОМОВ

4.1. АЛГОРИТМ МОРГАНА

Алгоритм Моргана для нумерации атомов в молекуле был предложен в 1965 г. Он основан на нумерации атомов по принципу *числа продленных связностей* (EC — *extended connectivity*).

Рассмотрим пошагово принцип нумерации согласно алгоритму Моргана.

Шаг 1. Поиск первого атома.

Сначала пронумеровывают все атомы по количеству связанных с ними атомов (n_1). Например, если атом связан с тремя атомами, то он имеет номер $n_1 = 3$, если с одним — то $n_1 = 1$ (рис. 38). Кратность связи в расчет не принимается. Атомы водорода можно не учитывать, так как их количество определяется исходя из правил валентности. По числу различных номеров атомов n_1 определяют параметр EC_0 , который равен числу различных n_1 .

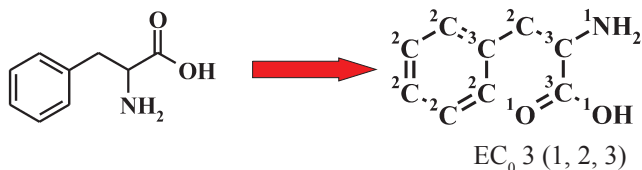


Рис. 38. Расчет n_1 и EC_0

Проводят расчет числа продленных связностей (EC) каждого атома, которое будет равно сумме всех соседних n_1 (рис. 39).

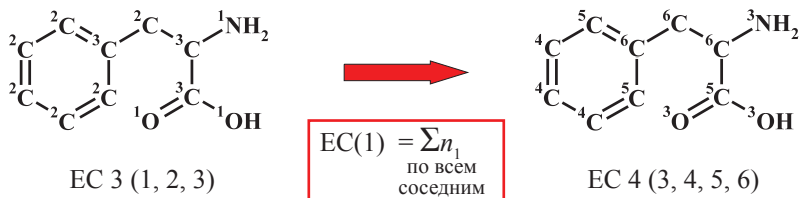


Рис. 39. Расчет EC_1

После сравнивают между собой все EC(1), и если среди них нет наибольшего, то проводят вторую итерацию расчета EC(2) (рис. 40).

$$EC(2) = \sum_{\substack{\text{по всем} \\ \text{соседним}}} EC(1)$$

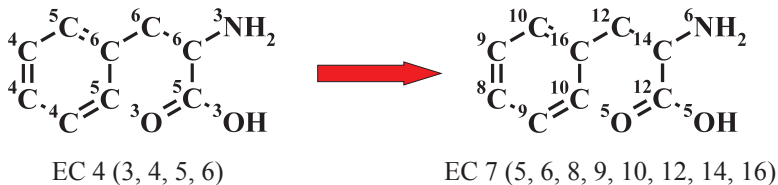


Рис. 40. Расчет EC(2) и EC₂

Этот шаг повторяют до тех пор, пока одно из значений EC(*i*) не получится больше других или пока количество продленных связей не станет постоянным (рис. 41).

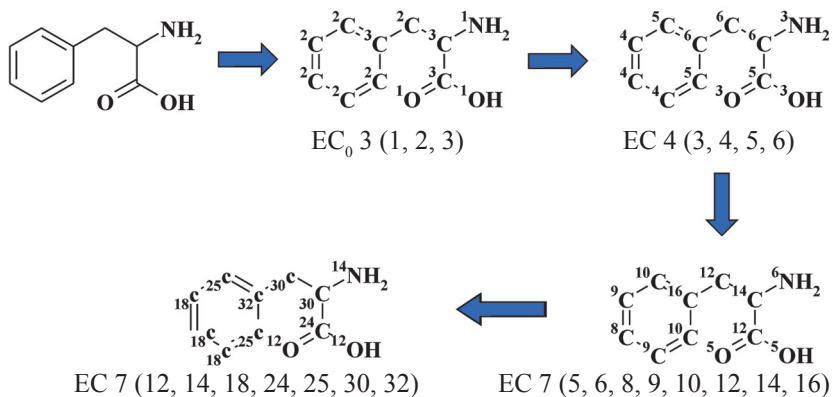


Рис. 41. Расчет EC_{*i*}

Шаг 2. Нумерация атомов.

Атому, имеющего наивысшее количество продленных связей, присваивают первый номер (рис. 42).

После нахождения атома с наивысшим ЕС проводят нумерацию всех атомов по принципу увеличения номера атома у атомов, ближайших к наименьшему номеру.

Атом, имеющий наивысшее значение ЕС, имеет номер **1**. Рассматривают все соединенные с ним атомы. Номер **2** получает атом с наибольшим ЕС, номера **3, 4** и т. д. получают все остальные атомы, соединенные с первым по уменьшению ЕС.



Рис. 42. Расчет номеров атомов по ЕС_{*i*}

Если два и более атомов, соединенных с первым, имеют одинаковое количество продленных связей ЕС, то нумерацию производят в соответствии с типом атома или связи.

По типу атома. Чем меньше номер атома, тем он младше, например, углерод (C) младше азота (N), который младше кислорода (O), который младше серы (S).

Если атомы одинаковы, то старше наименьший изотоп (¹²C старше ¹³C). Для одинаковых изотопов рассматривают абсолютную величину заряда: чем больше заряд атома, тем старше атом (O²⁻ младше O¹⁺).

По типу связи. Одинарная связь старше двойной, которая старше тройной. При равенстве связей принимают произвольное решение по старшинству атомов.

На рис. 43 приведен итоговый вид алгоритма Морган.

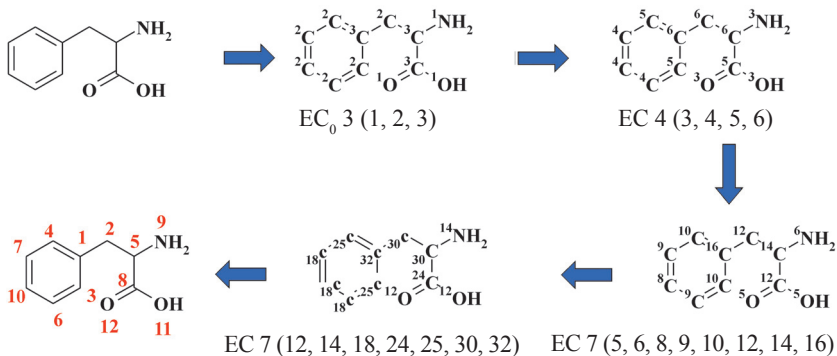


Рис. 43. Итоговый вид алгоритма Моргана

4.2. CANGEN-АЛГОРИТМ

Существуют некоторые отличия алгоритма CANGEN (CANonically GENerated) от алгоритма Моргана. Алгоритм Моргана рассматривает сумму близлежащих связей, CANGEN кроме этого учитывает также порядок связей, атомы, заряды и количество присоединенных атомов водорода.

Рассмотрим пошагово принцип нумерации согласно алгоритму CANGEN.

Шаг 1. Введение атомных инвариантов.

К атомным инвариантам относятся:

- 1) число соединений у атома;
- 2) число связей с тяжелыми атомами;
- 3) атомный номер;
- 4) знак заряда;
- 5) значение заряда по модулю;
- 6) количество атомов водорода.

В табл. 17 приведены инварианты наиболее распространенных групп и атомов.

Инварианты некоторых групп и атомов

Группа/атом	Инвариант						Запись
	1)	2)	3)	4)	5)	6)	
$-\text{CH}_2-$	2	2	6	0	0	2	20306002
$-\text{CH}_3$	1	1	6	0	0	3	10106003
$\begin{array}{c} \text{---C---} \\ \\ \text{O} \end{array}$	2	4	6	0	0	0	30406000
	1	2	8	0	0	0	10208000
$\begin{array}{c} \text{O}^- \\ \\ \text{N}^+ \\ \\ \text{O} \end{array}$	1	1	8	1	1	0	10108110
	3	4	7	2	1	0	30407210
	1	2	8	0	0	0	10208000

Шаг 2. Присвоение номера по ранжированию.

Рассмотрим на примере *n*-пентана:



Сначала молекула пентана преобразуется в вид:

10106003-20206002-20206002-20206002-1016003.

Далее присваиваются номера по рангу:

$-\text{CH}_3$ 10106003 → 1

$-\text{CH}_2-$ 20306002 → 2

1-2-2-2-1

Шаг 3. Суммирование всех соседних номеров.

Проводим суммирование всех соседних номеров и «молекула» преобразуется в вид:

1-2-2-2-1.

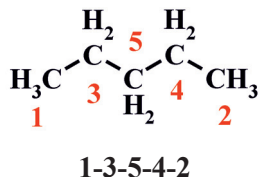
После суммирования получаем:

2-3-4-3-2.

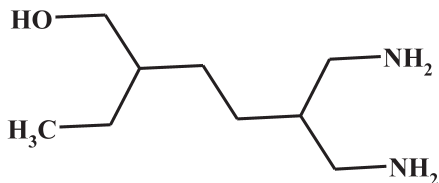
Проводим ранжирование аналогично шагу два. Номер два стал наименьшим — ему присваиваем значение 1, номеру 3 присваиваем значение 2, а номеру 4 — значение 3.

Запись **2-3-4-3-2** преобразуется в **1-2-3-2-1**.

Если все атомы получают свои номера или структура симметрична, то преобразуем нумерацию в конечную. Так, для симметричного случая запись **1-2-3-2-1** преобразуется в **1-3-5-4-2**.



Рассмотрим принцип нумерации согласно алгоритму CANGEN на примере еще одного соединения: 6-амино-2-этил-5-(аминоэтил)-1-гексанол.



Присваиваем каждому атому свой номер согласно инвариантам (рис. 44).

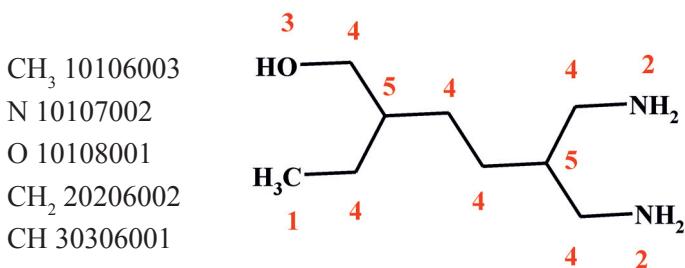


Рис. 44. Расчет инвариантов и номеров

И далее последовательно повторяем шаг 3 и шаг 2 (рис. 45).

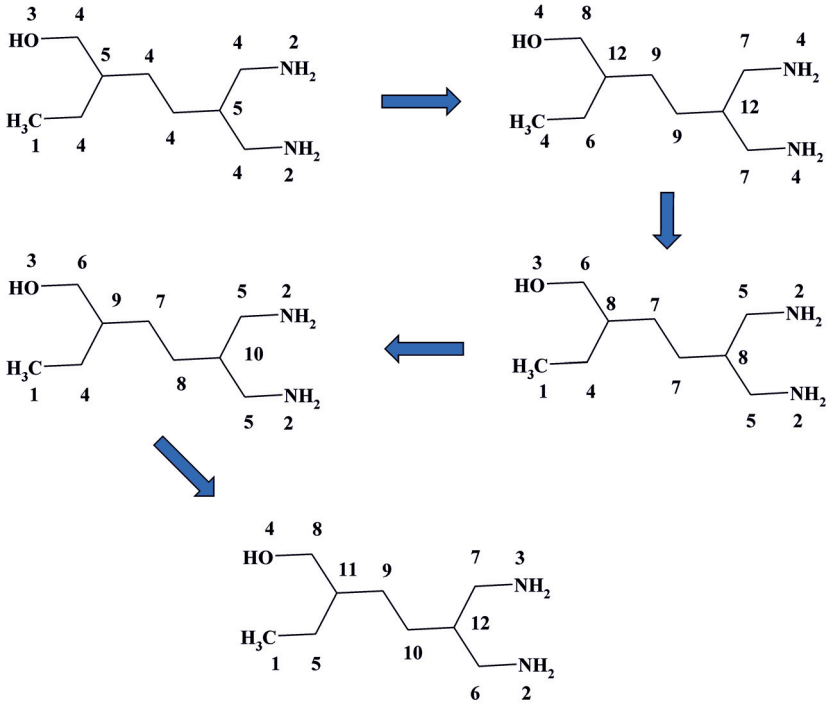
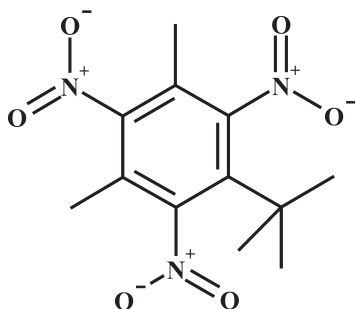


Рис. 45. Последовательное ранжирование и нумерация атомов

Таким образом, получаем окончательную нумерацию атомов в молекуле.

5. МЕЖДУНАРОДНЫЙ ХИМИЧЕСКИЙ ИДЕНТИФИКАТОР (INCHI)

Из-за наличия различных форм записи соединений для компьютерного анализа, как в виде линейных нотаций, так и в табличном виде, существуют проблемы идентификации структур. Речь идет от том, что различные поисковые системы используют разные способы записи молекул. Так, если одно соединение, например, дизамещенный тринитротолуол, мы запишем в нотации SMILES, MDL MOL и в виде таблиц связей, то получатся записи, содержащие, в зависимости от кодировки, различное количество бит информации (рис. 46).



SMILES: 62 bytes

O=[N+]([O-])c1c(c(c(c1C)[N+](=O)[O-])C(C)(C)C)[N+](=O)[O-]C

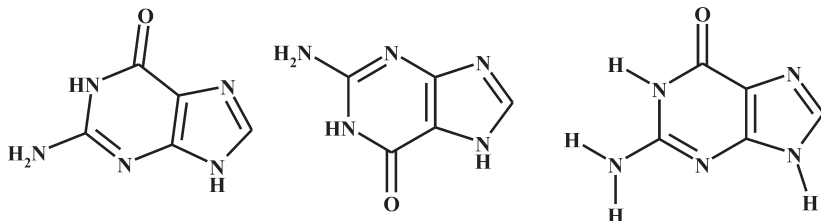
MDL MOL: 2066 bytes

Таблица связей: 998 bytes

Рис. 46. Различные записи тринитротолуола

Также в различных поисковых системах одно и то же соединение имеет свой (разный) персональный номер и соответственно может иметь неодинаковую форму записи в линейном виде для разных таутомеров и резонансных структур (рис. 47). Из-за существования таких разногласий в линейном описании молекул ИЮПАК решил

предложить новую систему записи на основе определенных требований к описанию молекул.

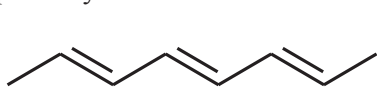


CAS Registry Number: 73-10-5 c1([nH]c(c2c(n1)[nH]cn2)=O)N
 Beilstein Registry Number: 9680 O=C1C2=C(NC=N2)N=C(N)N1
 Gmelin Registry Number: 431879 O=C1NC(N)=NC2=C1N=CN2
 MDL number: MFCD00071533 [nH]1c(nc2c(c1=O)nc[nH]2)N
 InChI=1/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4, 6, 7, 8, 9, 10, 11)

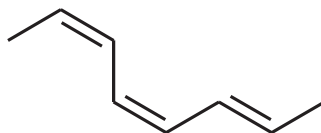
Рис. 47. Различные записи гуанидина

Можно сформулировать следующие формальные требования к описанию органических соединений для их компьютерного сравнения.

- Разные изомерные структуры должны быть записаны по-разному:



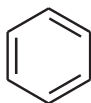
C/C=C/C=C/C=C/C



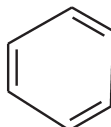
C/C=C\C=C/C=C/C/C

- Разные представления (резонансные структуры, таутомеры) одного и того же соединения должны быть записаны одинаково.

Примеры написания:

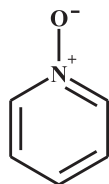


c1ccccc1

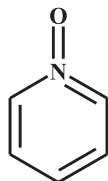


C1=CC=CC=C1

Бензол и циклогексаatriен

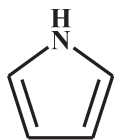


[O-][n+]1cccc1

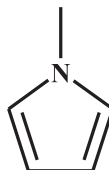


O=n1cccc1

Н-оксид

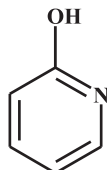


[nH]1cccc1

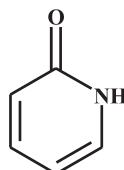


Cn1cccc1

Гомологи

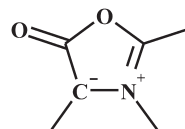
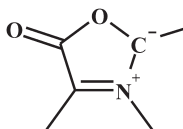
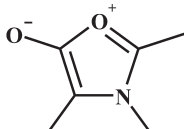
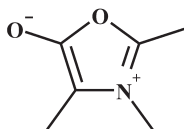


Oc1ncccc1



O=c1[nH]cccc1
O=CNC=CC=C1

Таутомеры



CC1=C([O-])OC(C)=[N+]1C
CC1=C([O-])[O+]=C(C)N1C
CC1=[N+](C)[C-](C)OC1=O
C[C-]([N+](C)=C(C)O1)C1=O

InChI=1/C6H9NO2/c1-4-6(8)9-5(2)7(4)3/h1-3H3

Резонансные структуры мезоионных гетероциклов

Чтобы избежать различий в написании молекул в компьютерном виде, в 2000–2005 гг. была разработана и апробирована на доступных базах данных молекул новая форма их записи: код *InChI* (*International Chemical Identifier* — международный химический идентификатор). Этот код представляет собой стандартизованную структурную текстовую линейную нотацию для обозначения молекул, обеспечения стандартного способа кодирования молекулярной информации и облегчения поиска такой информации в базах данных и в интернете. Нотация *InChI* была принята ИЮПАК в 2006 г. в качестве стандарта с открытым кодом.

Одновременно с кодом *InChI* была принята разработанная на его основе линейная нотация *InChIKey* — для использования ее в базах данных. Следует отметить, что нотация *InChIKey* была разработана для компьютерного поиска, но при этом она широко используется в научной литературе для обозначения и идентификации молекул, которые упоминаются в научных исследованиях.

5.1. ПРАВИЛА INCHI

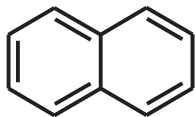
- Запись начинается с **InChI=**, затем следует номер версии 1.
- Далее следуют *слои* и *подслои*.

Слои:

1. Основной слой:
 - 1.1. Брутто-формула (без префикса);
 - 1.2. Связи, но не формальный порядок связи (префикс *c*);
 - 1.3. Атомы водорода (префикс *h*).
2. Слой изотопов (префикс *i*).
3. Слой зарядов:
 - 3.1. Слой положительных зарядов (префикс *p*);
 - 3.2. Слой отрицательных зарядов (префикс *q*).
4. Слой стереохимии (префикс *s* = 1 — абсолютная, 2 — относительная, 3 — рацемат):
 - 4.1. Двойные связи (Z/E, префикс *b* — *cis* –, *trans* +);
 - 4.2. Тетраэдрическая (sp^3 , префикс *t* — *S* –, *R* +).
5. Таутомеры (префикс *m*: 1 — да, 0 — нет).

• При записи *InChI* связи обозначаются тире (–), вне зависимости от кратности связи, разветвления обозначаются скобками ().

На рис. 48 приведены примеры записей *InChI* для нескольких молекул.



Основной слой

InChI=1/C10H8/c1-2-6-10-8-4-3-7-99(10)5-1/h1-8H

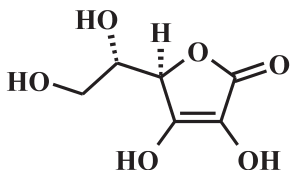
Химическая формула Связи между атомами
(блок связей начинается с c)

a

CH3CH2OH

InChI=1/C2H6O/c1-2-3/h3H, 2H2, 1H3

б



InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H, 1H2/t2-, 5+/m0/s1

в

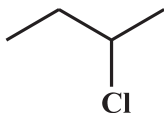
Рис. 48. Пример записи молекул:

a — нафталин; *б* — этанол; *в* — витамин С

5.1.1. Нумерация атомов (Color List)

Для записи связей между атомами в нотации *InChI* используется своя нумерация, которая носит название *список цветов (Color List)*.

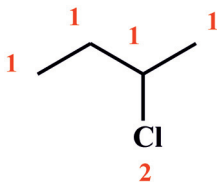
Нумерацию атомов рассмотрим на примере 2-хлорбутана.



Шаг 1. Сначала записывается брутто-формула:



Шаг 2. Атомам углерода **C** присваивается номер **1**. Атомы водорода не учитываются, остальные нумеруются согласно правилам записи брутто-формул, т. е. по алфавиту: атому **Cl** присваивается номер **2**. Получаем:



Шаг 3. Далее нумерация производится по количеству связей. После присвоения номеров по брутто-формуле через запятую к каждому номеру атома добавляется количество связей от данного атома:

— первый углеродный атом **C** связан с одним атомом → ему присваивают номер **1, 1**;

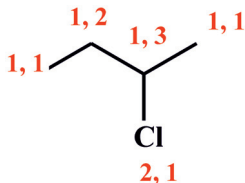
— второй **C**-атом связан с двумя атомами → номер **1, 2**;

— третий **C**-атом связан с тремя атомами → номер **1, 3**;

— четвертый **C**-атом связан с одним атомом → номер **1, 1**;

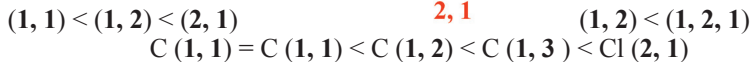
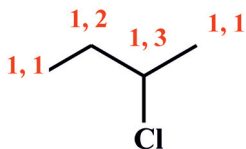
— атом хлора (**Cl**) связан с одним атомом → номер **2, 1**.

Получаем следующую нумерацию:



Шаг 4. Затем расставляем атомы в ряд ранжированием по принципу: **(1, 1) < (1, 2) < (2, 1); (1, 2) < (1, 2, 1)**.

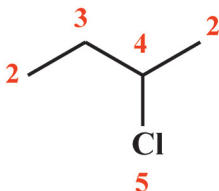
В результате получаем новую последовательность атомов, которым присваивают новые номера (цвета):



Присваиваем новые номера:



В результате получаем следующую нумерацию:

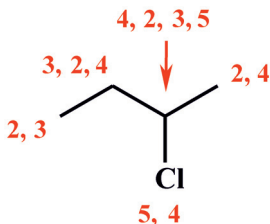


Шаг 5. Записываем к новым номерам (цветам) атомов через запятую номера атомов, с которыми они связаны:

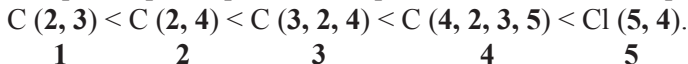
— углеродный атом $C(1, 1)$ стал у нас номером 2 , он связан с одним атомом — $C(1, 2)$, который стал номером $3 \rightarrow$ номер $2, 3$;
 — второй C -атом 3 связан с двумя атомами — 2 и $4 \rightarrow$ номер $3, 2, 4$;

— третий C -атом связан с тремя атомами \rightarrow номер $4, 2, 3, 5$;
 — четвертый C -атом связан с одним атомом \rightarrow номер $2, 4$;
 — атом хлора (Cl) связан с одним атомом \rightarrow номер $5, 4$.

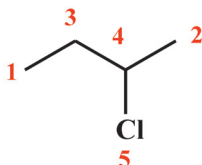
Получаем:



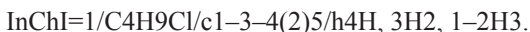
Шаг 6. Затем повторяем действия четвертого шага: расставляем атомы в ряд по ранжированию и присваиваем новые номера (цвета):



Нумерацию атомов заканчиваем, когда каждому атому присвоен индивидуальный номер (цвет). Таким образом, получаем:



При записи *InChI* связи обозначаются тире (–), вне зависимости от кратности связи, разветвления обозначаются скобками ().



Пример нумерации для фенилаланина представлен на рис. 49.

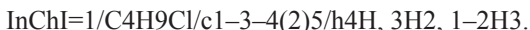
5.1.2. Записи основных слоев

Запись расположения атомов водорода обозначается отдельно, после указания связей и разветвлений.

Порядок записи атомов водорода

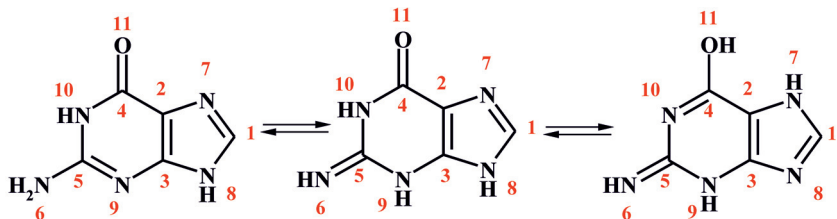
Сначала пишется /h, затем номера (цвета) атомов с одним водородом и буква **H**, далее номера (цвета) атомов с двумя водородами, затем **H2**, потом номера (цвета) атомов с тремя водородами, и **H3**.

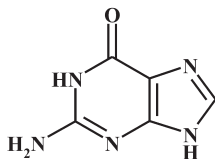
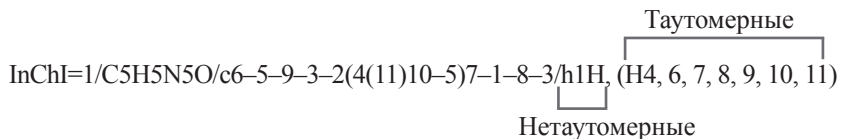
Пример записи для 2-хлорбутана:



Обозначение таутомеров

После того как произведена нумерация атомов в молекуле, записаны брутто-формула, связи, положения водородов, в скобках указываются все возможные положения таутомерных водородов для всех таутомерных форм соединения (рис. 50).



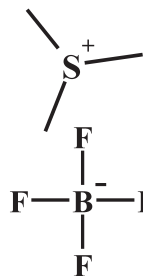


/h1H, (H4, 6, 7, 8, 9, 10, 11)

Рис. 50. Запись таутомеров в InChI

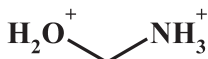
Слой зарядов

Для отрицательных зарядов запись начинается с буквы **q**, для положительных зарядов — с буквы **p**. Однако записывается не распределение зарядов в молекуле, а только общий заряд молекулы.



InChI=1/C3H9S/c1-4(2)3/h1-3H3/p+1

InChI=1/BF4/c2-1(3, 4)5/q-1

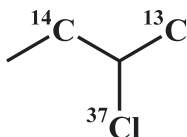


InChI=1/CH5NO/c2-1-3/h3H, 1-2H2/p+2

Слой изотопов

Запись изотопов проводят аналогично записи зарядов, начиная с буквы **i**. Вначале указывается номер атома, затем отклонение по массе от наиболее распространенного изомера:

- 1, iso_weight(1),
- 2, iso_weight(2),
- n, iso_weight(n)



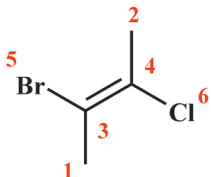
Так, для 2-хлорбутана, имеющего изотопный атом углерода, запись будет иметь вид:



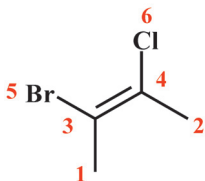
Слой стереохимии

Сначала рассматривают двойные связи типа $>X=Y<$ и кумулены типа $>W=X=Y=Z<$.

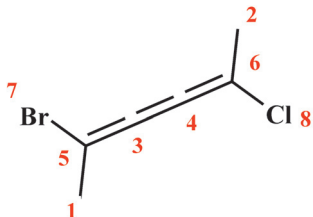
Префикс *b*, *cis*-положение отмечают знаком «-», *trans*-положение отмечают знаком «+».



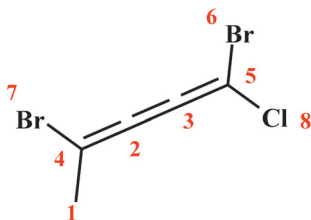
trans-положение $\text{InChI}=1/\text{C}4\text{H}6\text{BrCl}/\text{c}1-3(5)4(2)6/\text{h}1-2\text{H}3/\text{b}4-3+$



cis-положение $\text{InChI}=1/\text{C}4\text{H}6\text{BrCl}/\text{c}1-3(5)4(2)6/\text{h}1-2\text{H}3/\text{b}4-3-$



$\text{InChI}=1/\text{C}6\text{H}6\text{BrCl}/\text{c}1-5(7)3-4-6(2)8/\text{h}1-2\text{H}3/\text{b}6-5+$



InChI=1/C5H3Br2Cl/c1-4(6)2-3-5(7)8/h1H3/b5-4+

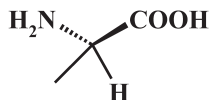
Далее записывают sp^3 -гибридизованные атомы и кумулены типа $>X=Y=Z<$.

Префиксом *t* со знаком «±» обозначают следующую конфигурацию: *S* отмечается знаком «-», *R* отмечается знаком «+».

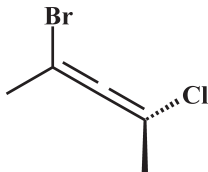
Префиксом *s* обозначается абсолютная конфигурация (**s1**), относительная (**s2**), рацемат (**s3**).



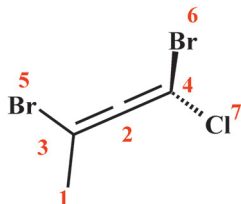
InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H, 4H2, 1H3, (H,5,6)/t2-/m0/s1



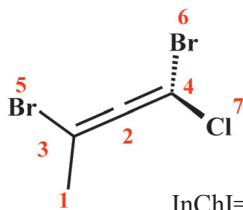
InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H, 4H2, 1H3, (H, 5, 6)/t2+/m0/s1



InChI=1/C5H6BrCl/c1-4(6)3-5(2)7/h1-2H3/t3-/m0/s1

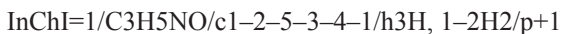
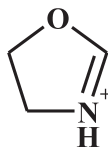
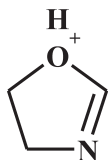


InChI=1/C4H3Br2Cl/c1-3(5)2-4(6)7/h1H3/t2-/m0/s1

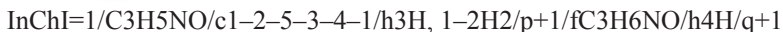
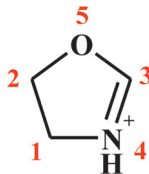
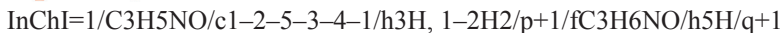
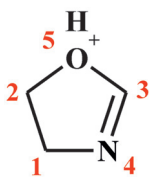


Слой фиксированных Н

Слой необходим для того, чтобы выделить конкретный таутомер, а также показать место протонирования соединения. В общем виде запись без указания таутомеров будет выглядеть следующим образом:

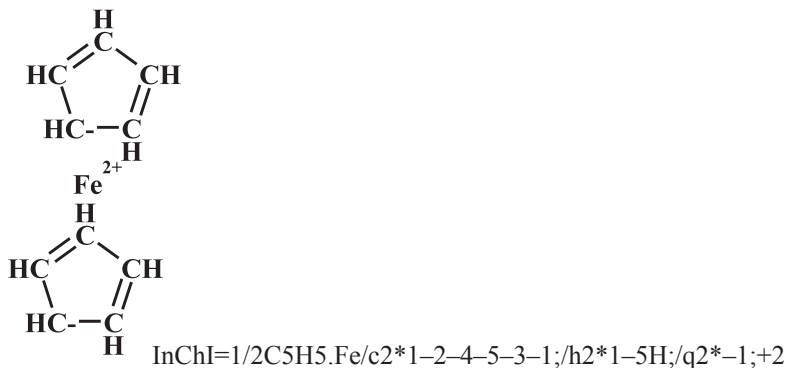


Для разных таутомеров к записи кода *InChI* добавляется слой фиксированных водородов с буквы *f* и указанием брутто-формулы протонированной формы молекулы, затем указывается номер атома с фиксированным водородом.



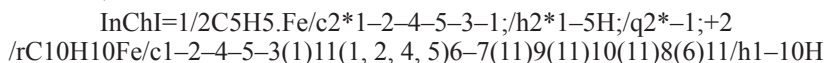
Нековалентно связанные атомы

Запись кода *InChI* для нековалентно связанных атомов будет следующая: в основном слое записываем нековалентно связанные части молекул как самостоятельные молекулы (их в брутто-формуле перечисляют через точку), нумерацию атомов проводим для каждой части нековалентно связанных молекул, цвета (номера) атомов отдельных частей записываем через точку с запятой (;), аналогично записываем и слой водородов, а также слой зарядов.



В дополнительном слое фиксированных связей, начиная с буквы *r*, записывают общую брутто-формулу молекулы, все связи рисуют как ковалентные и проводят расчет цветов атомов как для одной молекулы, соответственно записывают слой связей и слой водородов с использованием общего списка цветов (*Color List*).

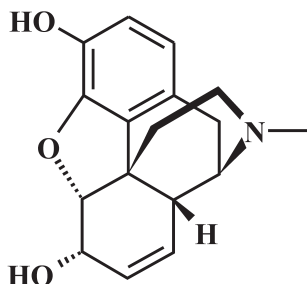
Так, для нижеприведенного соединения, имеющего связи с комплексообразователем, запись с помощью нотации *InChI* выглядит следующим образом:



5.2. INCHIKEY — КЛЮЧ ДЛЯ ПОИСКА СТРУКТУРЫ

Следует отметить, что несмотря на унификацию линейного кода *InChI*, которая позволяет одинаково легко записывать как резонансные структуры, так и таутомерные, проблема записи соединения для баз данных осталась нерешенной, поскольку запись даже одного и того же соединения имеет различную длину в зависимости от рисунка структурной формулы. Для баз данных желательна запись фиксированной длины. С этой целью был предложен код записи *InChIKey*, который имеет фиксированное количество знаков. Преобразование записи *InChI* в формат *InChIKey* возможно с помощью *Secure Hasg Algorithm 2* — алгоритма криптографического хеширования.

Код *InChI* для такого сложного соединения, как морфин, выглядит довольно громоздко:



InChI=1/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11, 13, 16, 19-20H, 6-8H2, 1H3/t10-, 11-,13-,16-,17-/m0/s1

Запись в формате *InChIKey* для той же самой структуры будет выглядеть следующим образом:

BQJCRHNNABKAKU-XKQOQXLYSA-N.

Вообще хеш-функции предназначены для создания «отпечатков» или «дайджестов» сообщений произвольной битовой длины и применяются в различных приложениях или компонентах, связанных с защитой информации.

Хеширование (*hashing*) — преобразование входного массива данных произвольной длины в выходную битовую строку фиксированной длины. Исходное сообщение после дополнения разбивается на блоки, каждый блок — на 8 слов. Алгоритм пропускает каждый блок сообщения через цикл с 64 или 80 итерациями (раундами). На каждой итерации 2 слова из восьми преобразуются, функцию преобразования задают остальные слова. Результаты обработки каждого блока складываются, сумма является значением хеш-функции.

Было принято, что *InChIKey* содержит 25 фиксированных знаков. Первые 14 знаков кодируют молекулярный скелет, следующие 8 знаков — другие слои (рис. 51), затем следует двухбуквенное обозначение версии записи (*SA*) и последний знак, который характеризует заряд всей молекулы (табл. 18).

Блок молекулярного скелета

Блок стереохимии, изотопов, зарядов

Знак числа протонов:
N — нейтральный характер;
M — для -1 водород;
O — для +1 водород
и т. д.

InChIKey = BQJCRHNNABKAKU-XKUOQXLYSA-N

Знак S — стандартный InChIKey

Знак: A — версия 1;

B — версия 2

Рис. 51. Состав записи *InChIKey*

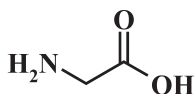
Таблица 18

Буквенное обозначение заряда молекулы в *InChIKey*

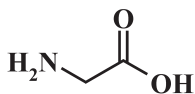
Заряд	Протоны	Заряд	Протоны
N	0		
M	-1	O	+1
L	-2	P	+2
K	-3	Q	+3
J	-4	R	+4
I	-5	S	+5
H	-6	T	+6

Заряд	Протоны	Заряд	Протоны
G	-7	U	+7
F	-8	V	+8
E	-9	W	+9
D	-10	X	+10
C	-11	Y	+11
B	-12	Z	+12
A	< -12 или > +12		

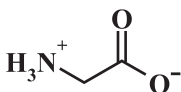
Примеры кодирования в формате *InChIKey* представлены ниже:



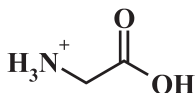
InChI = 1S/C2H5NO2/c3-1-2(4)5/h1, 3H2, (H, 4, 5)
InChIKey = DHMQDGOQFOQNFH-UHFFFAOYSA-N



InChI = 1S/C2H5NO2/c3-1-2(4)5/h1, 3H2, (H, 4, 5)/p-1
InChIKey = DHMQDGOQFOQNFH-UHFFFAOYSA-M



InChI = 1S/C2H5NO2/c3-1-2(4)5/h1, 3H2, (H, 4, 5)
InChIKey = DHMQDGOQFOQNFH-UHFFFAOYSA-N



InChI=1S/C2H5NO2/c3-1-2(4)5/h1, 3H2, (H, 4, 5)/p+1
InChIKey= DHMQDGOQFOQNFH-UHFFFAOYSA-O

Информационный поиск по запросу, состоящему из кода *InChI* или кода *InChIKey*, пока что не обладает какими-то особыми преимуществами по сравнению с другими методами. Причина кроется в том, что эти коды — явление новое, в настоящее время оно не получило повсеместного распространения несмотря на то, что эти коды приняты ИЮПАК в качестве обязательных правил номенклатуры. Тем не менее есть все основания полагать, что такая

ситуация временна. Научному сообществу необходима стандартная линейная нотация химической структуры, и *InChIKey* была создана для этой роли.

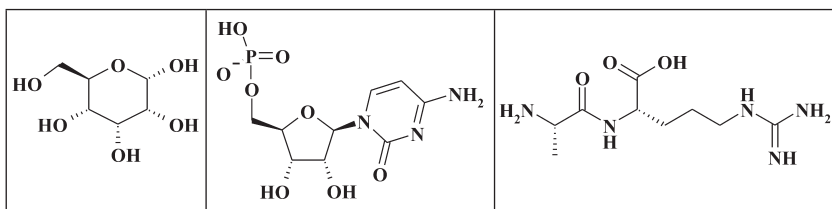
В интернете растет количество ресурсов, благодаря которым можно узнать код *InChI* для химического вещества. Большинство научных систем поиска структур соединений, баз данных, химических журналов начинают использовать данную нотацию.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

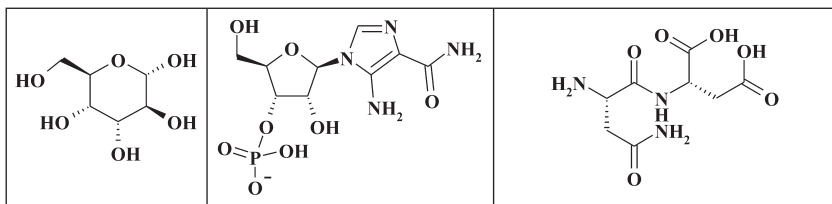
Задание 1

Пронумеровать молекулы, используя алгоритм Моргана, и записать в WLN, ROSDAL, SMILES и SLN.

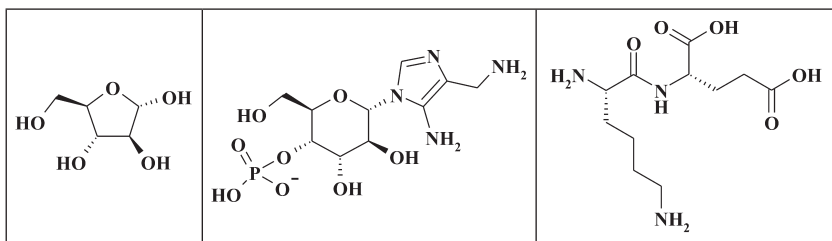
ВАРИАНТ 1



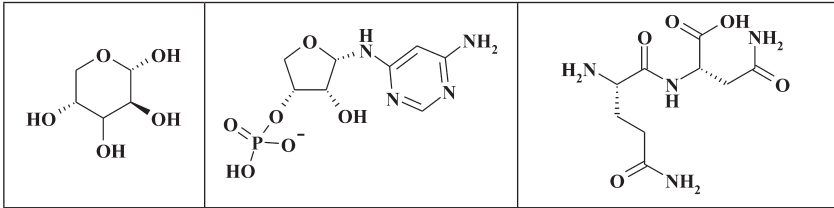
ВАРИАНТ 2



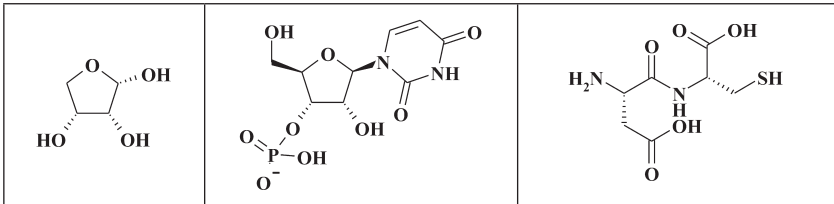
ВАРИАНТ 3



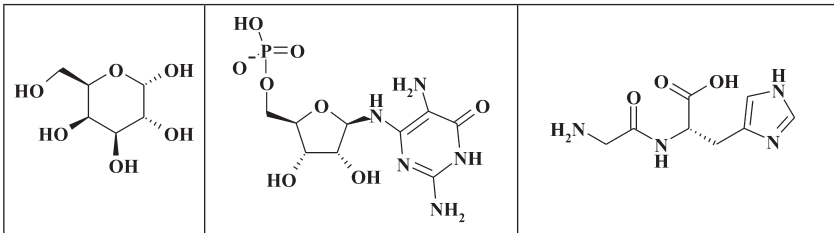
ВАРИАНТ 4



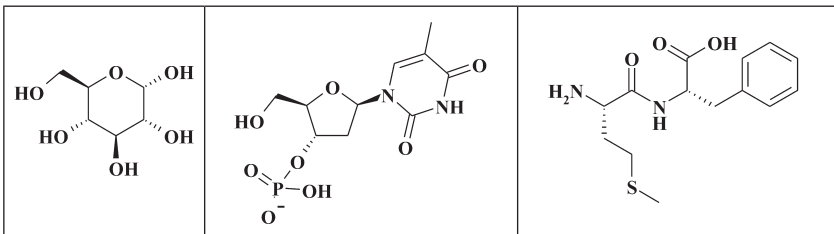
ВАРИАНТ 5



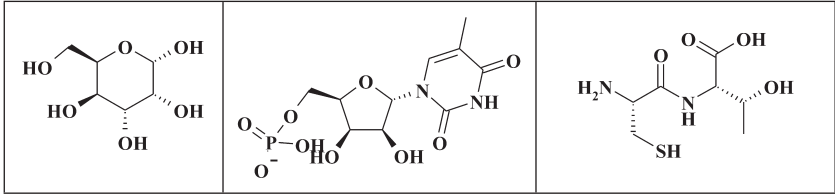
ВАРИАНТ 6



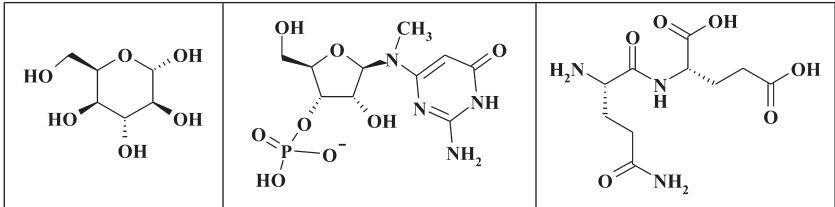
ВАРИАНТ 7



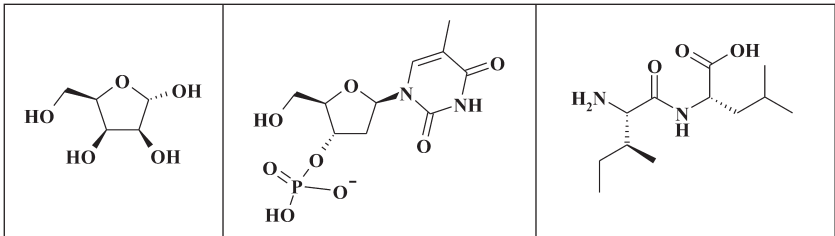
ВАРИАНТ 8



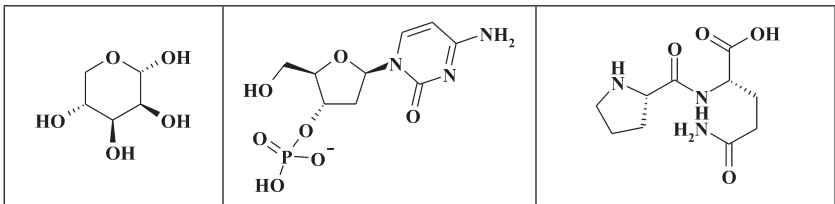
ВАРИАНТ 9



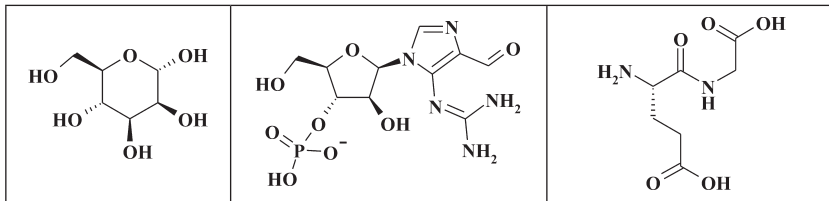
ВАРИАНТ 10



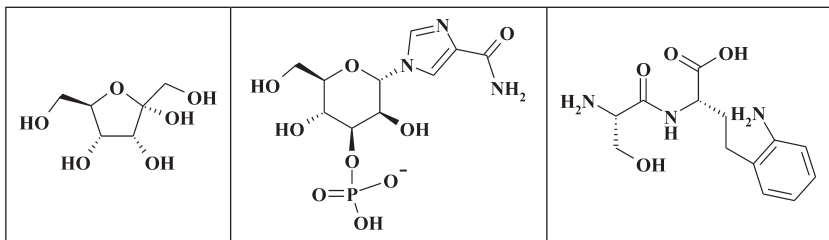
ВАРИАНТ 11



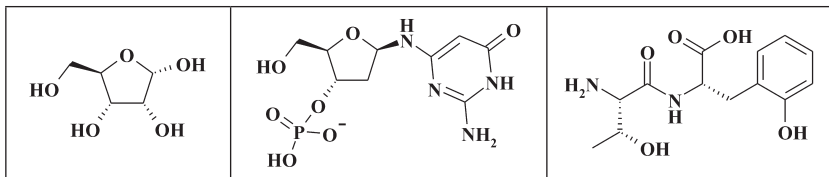
ВАРИАНТ 12



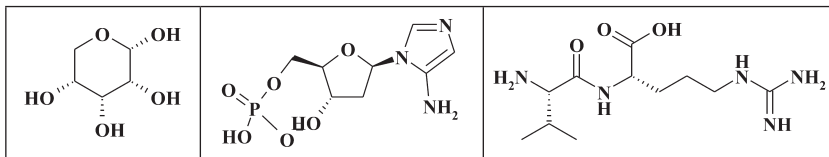
ВАРИАНТ 13



ВАРИАНТ 14



ВАРИАНТ 15



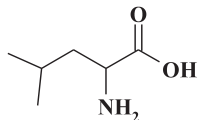
Задание 2

ВАРИАНТ 1

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в системе WLN, SMILES и SLN.

```
-ISIS- 10130513072D
11 11 0 0 0 0 0 0 0 0999 V2000
  -3.2893 1.1340 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  -3.2815 1.9375 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  -2.4962 2.2135 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -2.4962 3.0410 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  -3.9663 2.4051 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -3.9045 3.2340 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -4.5885 3.7015 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -5.3369 3.3423 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -5.3971 2.5109 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -4.7122 2.0472 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -1.8231 1.7366 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  5 6 2 0 0 0 0
  1 2 1 0 0 0 0
  6 7 1 0 0 0 0
  2 3 1 0 0 0 0
  7 8 2 0 0 0 0
  3 4 2 0 0 0 0
  8 9 1 0 0 0 0
  2 5 1 0 0 0 0
  9 10 2 0 0 0 0
  10 5 1 0 0 0 0
  3 11 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.469900 1 0.000000 0 0.000000 0 1 0 0
```

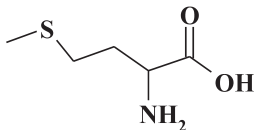
C	1.519900	1	109.470900	1	0.000000	0	2	1	0
O	1.220000	1	119.999900	1	179.999917	1	3	2	1
C	1.540000	1	109.471000	1	240.000500	1	2	1	3
C	1.519900	1	109.471000	1	180.000003	1	5	2	1
O	1.219900	1	120.000000	1	179.999953	1	6	5	2
N	1.320000	1	119.999900	1	359.999953	1	6	5	2
O	1.499900	1	120.000000	1	359.999917	1	3	2	1
H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

ВАРИАНТ 2

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513072D
9 8 0 0 0 0 0 0 0 0999 v2000
 2.7490  1.0140  0.0000  N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.7473  1.9020  0.0000  C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
 3.4773  2.2688  0.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3.4773  3.0797  0.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.0216  2.2735  0.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1.3430  1.8280  0.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.5852  2.1933  0.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1.3759  1.0411  0.0000  N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.1795  1.8357  0.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
2 5 1 0 0 0 0
6 7 2 0 0 0 0
6 8 1 0 0 0 0
5 6 1 0 0 0 0
3 9 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000  0 0.000000  0 0.000000  0 0 0 0
C 1.469900  1 0.000000  0 0.000000  0 1 0 0
C 1.519900  1 109.470900  1 0.000000  0 2 1 0
O 1.220000  1 119.999900  1 179.999917  1 3 2 1
C 1.540000  1 109.471000  1 240.000500  1 2 1 3
C 1.519900  1 109.471000  1 180.000003  1 5 2 1
O 1.219900  1 120.000000  1 179.999953  1 6 5 2
N 1.320000  1 119.999900  1 359.999953  1 6 5 2
O 1.499900  1 120.000000  1 359.999917  1 3 2 1
```

H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

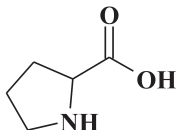
ВАРИАНТ 3

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513072D
10 9 0 0 0 0 0 0 0 0999 v2000
  8.1532  1.1290  0.0000  N  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  8.1279  1.9770  0.0000  C  0  0  3  0  0  0  0  0  0  0  0  0  0
  8.8914  2.2510  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0  0
  8.8914  3.0648  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0  0
  7.4514  2.4373  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0  0
  6.6791  2.1601  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0  0
  6.1644  2.7953  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0  0
  6.4497  3.5835  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0  0
  5.3544  2.6584  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0  0
  9.5662  1.7763  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0  0
2  1  1  0  0  0  0
2  3  1  0  0  0  0
3  4  2  0  0  0  0
2  5  1  0  0  0  0
6  5  1  0  0  0  0
7  8  2  0  0  0  0
6  7  1  0  0  0  0
7  9  1  0  0  0  0
3 10  1  0  0  0  0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N  0.000000  0  0.000000  0  0.000000  0  0  0  0
C  1.445999  1  0.000000  0  0.000000  0  1  0  0
C  1.540009  1  109.500000  1  0.000000  0  2  1  0
C  1.540009  1  109.500000  1  60.000000  1  3  2  1
O  3.531514  1  29.907623  1  225.454373  1  1  2  3
C  1.540009  1  109.500000  1  120.070564  1  2  1  3
C  1.540009  1  109.500000  1  299.929436  1  3  2  1
O  1.410004  1  109.500000  1  300.000000  1  6  2  1
```

H	1.027985	1	109.500000	1	180.000000	1	1	2	3
H	1.027985	1	109.500000	1	59.929436	1	1	2	3
H	1.121994	1	109.500000	1	59.929436	1	2	1	9
H	1.121994	1	109.500000	1	180.070564	1	3	2	1
H	1.121994	1	109.500000	1	179.999999	1	4	3	2
H	1.121994	1	109.500000	1	59.929436	1	4	3	2
H	1.121994	1	109.500000	1	300.070564	1	4	3	2
H	1.121994	1	109.500000	1	180.000000	1	7	3	2
H	1.121994	1	109.500000	1	59.929436	1	7	3	2
H	1.121994	1	109.500000	1	300.070564	1	7	3	2
H	0.991989	1	109.500000	1	180.000000	1	8	6	2

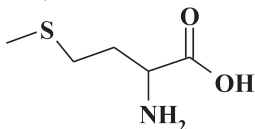
ВАРИАНТ 4

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513072D
9 8 0 0 0 0 0 0 0 0999 v2000
  2.7490  1.0140  0.0000  N  0  0  0  0  0  0  0  0  0  0  0  0
  2.7473  1.9020  0.0000  C  0  0  3  0  0  0  0  0  0  0  0  0
  3.4773  2.2688  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0
  3.4773  3.0797  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0
  2.0216  2.2735  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0
  1.3430  1.8280  0.0000  C  0  0  0  0  0  0  0  0  0  0  0  0
  0.5852  2.1933  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0
  1.3759  1.0411  0.0000  N  0  0  0  0  0  0  0  0  0  0  0  0
  4.1795  1.8357  0.0000  O  0  0  0  0  0  0  0  0  0  0  0  0
2 1 1 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
2 5 1 0 0 0 0
6 7 2 0 0 0 0
6 8 1 0 0 0 0
5 6 1 0 0 0 0
3 9 1 0 0 0 0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N  0.000000  0  0.000000  0  0.000000  0  0  0  0
C  1.469900  1  0.000000  0  0.000000  0  1  0  0
C  1.519900  1  109.470900  1  0.000000  0  2  1  0
O  1.220000  1  119.999900  1  179.999917  1  3  2  1
C  1.540000  1  109.471000  1  240.000500  1  2  1  3
C  1.519900  1  109.471000  1  180.000003  1  5  2  1
O  1.219900  1  120.000000  1  179.999953  1  6  5  2
N  1.320000  1  119.999900  1  359.999953  1  6  5  2
O  1.499900  1  120.000000  1  359.999917  1  3  2  1
```

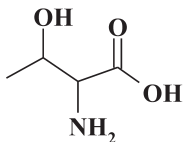

H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

ВАРИАНТ 5

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513072D
9 8 0 0 0 0 0 0 0 0999 V2000
-2.6843 -3.1735 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.6777 -2.2772 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.9436 -1.9104 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.9436 -1.0953 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-3.4075 -1.9057 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-4.0904 -2.3553 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-4.8523 -1.9858 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-4.0949 -3.1255 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.2418 -2.3442 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
2 5 1 0 0 0 0
6 7 2 0 0 0 0
6 8 1 0 0 0 0
5 6 1 0 0 0 0
3 9 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
H 0.960000 1 109.471000 1 180.000000 1 9 3 2
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.469900 1 0.000000 0 0.000000 0 1 0 0
C 1.519900 1 109.470900 1 0.000000 0 2 1 0
O 1.220000 1 119.999900 1 179.999917 1 3 2 1
C 1.540000 1 109.471000 1 240.000500 1 2 1 3
C 1.519900 1 109.471000 1 180.000003 1 5 2 1
O 1.219900 1 120.000000 1 179.999953 1 6 5 2
```

N	1.320000	1	119.999900	1	359.999953	1	6	5	2
O	1.499900	1	120.000000	1	359.999917	1	3	2	1
H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

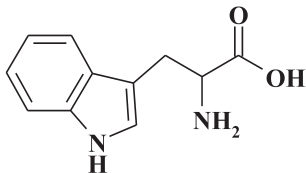
ВАРИАНТ 6

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513082D
7 6 0 0 0 0 0 0 0 0999 v2000
  1.7025 -3.1867 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  1.6771 -2.3292 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  2.4540 -2.0575 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  2.4540 -1.2383 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  0.9964 -1.8699 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  0.3376 -2.3510 0.0000 S 0 0 0 0 0 0 0 0 0 0 0 0
  3.1268 -2.5350 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
5 6 1 0 0 0 0
2 5 1 0 0 0 0
3 7 1 0 0 0 0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.445999 1 0.000000 0 0.000000 0 1 0 0
C 1.540009 1 109.500000 1 0.000000 0 2 1 0
C 1.540009 1 109.500000 1 60.000000 1 3 2 1
O 3.531514 1 29.907623 1 225.454373 1 1 2 3
C 1.540009 1 109.500000 1 120.070564 1 2 1 3
C 1.540009 1 109.500000 1 299.929436 1 3 2 1
O 1.410004 1 109.500000 1 300.000000 1 6 2 1
H 1.027985 1 109.500000 1 180.000000 1 1 2 3
H 1.027985 1 109.500000 1 59.929436 1 1 2 3
H 1.121994 1 109.500000 1 59.929436 1 2 1 9
```

H	1.121994	1	109.500000	1	180.070564	1	3	2	1
H	1.121994	1	109.500000	1	179.999999	1	4	3	2
H	1.121994	1	109.500000	1	59.929436	1	4	3	2
H	1.121994	1	109.500000	1	300.070564	1	4	3	2
H	1.121994	1	109.500000	1	180.000000	1	7	3	2
H	1.121994	1	109.500000	1	59.929436	1	7	3	2
H	1.121994	1	109.500000	1	300.070564	1	7	3	2
H	0.991989	1	109.500000	1	180.000000	1	8	6	2

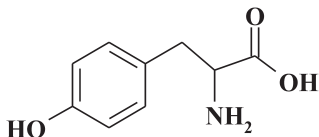
ВАРИАНТ 7

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513082D
11 11 0 0 0 0 0 0 0 0999 V2000
  1.7025 -3.1867 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  8.0802 -2.4911 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  8.0964 -1.6171 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  8.8693 -1.3456 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  8.8693 -0.5307 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  7.4241 -1.1577 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  6.7654 -1.6431 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.9763 -1.3921 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  6.7653 -2.4659 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.4974 -2.0794 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.9786 -2.7277 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  9.5424 -1.8225 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
2   1   1   0   0   0   0
2   3   1   0   0   0   0
3   4   2   0   0   0   0
2   5   1   0   0   0   0
6   7   1   0   0   0   0
6   8   2   0   0   0   0
7   9   2   0   0   0   0
8  10   1   0   0   0   0
9  10   1   0   0   0   0
5   6   1   0   0   0   0
3  11   1   0   0   0   0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.445999 1 0.000000 0 0.000000 0 1 0 0
O 3.100295 1 48.053010 1 0.000000 0 1 2 0
```

C	1.540009	1	109.500000	1	26.970371	1	2	1	3
O	1.410004	1	109.500000	1	60.000000	1	4	2	1
H	1.027985	1	109.500000	1	180.000000	1	1	2	4
H	1.027985	1	109.500000	1	59.929436	1	1	2	4
H	1.121994	1	109.500000	1	59.929436	1	2	1	6
H	1.121994	1	109.500000	1	300.070564	1	2	1	6
H	0.991989	1	109.500000	1	179.999999	1	5	4	2

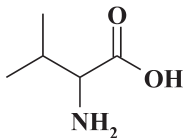
ВАРИАНТ 8

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513082D
9 8 0 0 0 0 0 0 0 0999 v2000
 13.3245 -2.8292 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 13.3245 -2.0017 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
 12.6034 -1.5859 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
 11.8904 -2.0017 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 14.0456 -0.7667 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 14.0456 -1.5858 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 11.1378 -1.6628 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 12.6033 -0.7667 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 14.7599 -1.9985 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 3 1 0 0 0 0
3 4 1 0 0 0 0
5 6 2 0 0 0 0
6 2 1 0 0 0 0
2 1 1 0 0 0 0
4 7 1 0 0 0 0
3 8 1 0 0 0 0
6 9 1 0 0 0 0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.469900 1 0.000000 0 0.000000 0 1 0 0
C 1.519900 1 109.470900 1 0.000000 0 2 1 0
O 1.220000 1 119.999900 1 179.999917 1 3 2 1
C 1.540000 1 109.471000 1 240.000500 1 2 1 3
C 1.519900 1 109.471000 1 180.000003 1 5 2 1
O 1.219900 1 120.000000 1 179.999953 1 6 5 2
```

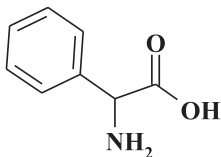

N	1.320000	1	119.999900	1	359.999953	1	6	5	2
O	1.499900	1	120.000000	1	359.999917	1	3	2	1
H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

ВАРИАНТ 9

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513082D
9 8 0 0 0 0 0 0 0 0999 v2000
  -2.7798 -8.2250 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  -2.7798 -7.4017 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  -3.4967 -6.9900 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -4.2054 -7.4016 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  -2.0628 -6.1708 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  -2.0628 -6.9899 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -4.9539 -7.0667 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -4.2958 -8.2247 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  -1.3488 -7.4032 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
2 3 1 0 0 0 0
3 4 1 0 0 0 0
5 6 2 0 0 0 0
6 2 1 0 0 0 0
2 1 1 0 0 0 0
4 7 1 0 0 0 0
8 4 1 0 0 0 0
6 9 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.445999 1 0.000000 0 0.000000 0 1 0 0
C 1.540009 1 109.500000 1 0.000000 0 2 1 0
C 1.540009 1 109.500000 1 60.000000 1 3 2 1
O 3.531514 1 29.907623 1 225.454373 1 1 2 3
C 1.540009 1 109.500000 1 120.070564 1 2 1 3
C 1.540009 1 109.500000 1 299.929436 1 3 2 1
O 1.410004 1 109.500000 1 300.000000 1 6 2 1
```

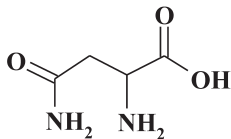
H	1.027985	1	109.500000	1	180.000000	1	1	2	3
H	1.027985	1	109.500000	1	59.929436	1	1	2	3
H	1.121994	1	109.500000	1	59.929436	1	2	1	9
H	1.121994	1	109.500000	1	180.070564	1	3	2	1
H	1.121994	1	109.500000	1	179.999999	1	4	3	2
H	1.121994	1	109.500000	1	59.929436	1	4	3	2
H	1.121994	1	109.500000	1	300.070564	1	4	3	2
H	1.121994	1	109.500000	1	180.000000	1	7	3	2
H	1.121994	1	109.500000	1	59.929436	1	7	3	2
H	1.121994	1	109.500000	1	300.070564	1	7	3	2
H	0.991989	1	109.500000	1	180.000000	1	8	6	2

ВАРИАНТ 10

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513082D
9 8 0 0 0 0 0 0 0 0999 v2000
  2.4731 -8.1803 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  2.4936 -7.3700 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  3.2737 -7.0959 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  3.2737 -6.2738 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  1.8047 -6.9055 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  1.0199 -7.1870 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  0.4968 -6.5433 0.0000 S 0 0 0 0 0 0 0 0 0 0 0 0
 -0.2862 -6.8211 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  3.9468 -7.5729 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 0 0 0 0
2 3 1 0 0 0 0
3 4 2 0 0 0 0
2 5 1 0 0 0 0
6 5 1 0 0 0 0
7 8 1 0 0 0 0
6 7 1 0 0 0 0
3 9 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.469900 1 0.000000 0 0.000000 0 1 0 0
C 1.519900 1 109.470900 1 0.000000 0 2 1 0
O 1.220000 1 119.999900 1 179.999917 1 3 2 1
C 1.540000 1 109.471000 1 240.000500 1 2 1 3
C 1.519900 1 109.471000 1 180.000003 1 5 2 1
O 1.219900 1 120.000000 1 179.999953 1 6 5 2
```

N	1.320000	1	119.999900	1	359.999953	1	6	5	2
O	1.499900	1	120.000000	1	359.999917	1	3	2	1
H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

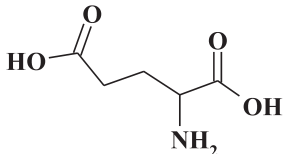
ВАРИАНТ 11

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513092D
8 8 0 0 0 0 0 0 0 0999 v2000
  7.0791 -7.3155 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
  7.1423 -6.4977 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0
  6.2746 -7.4934 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  6.3891 -6.1682 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  5.8613 -6.7714 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  7.8471 -6.0775 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
  7.8471 -5.2275 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
  8.5644 -6.4850 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 0 0 0 0
1 3 1 0 0 0 0
2 4 1 0 0 0 0
3 5 1 0 0 0 0
4 5 1 0 0 0 0
6 7 2 0 0 0 0
2 6 1 0 0 0 0
6 8 1 0 0 0 0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

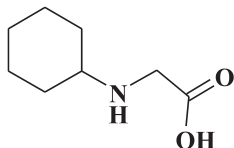
```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.445999 1 0.000000 0 0.000000 0 1 0 0
O 3.100295 1 48.053010 1 0.000000 0 1 2 0
C 1.540009 1 109.500000 1 26.970371 1 2 1 3
O 1.410004 1 109.500000 1 60.000000 1 4 2 1
H 1.027985 1 109.500000 1 180.000000 1 1 2 4
H 1.027985 1 109.500000 1 59.929436 1 1 2 4
H 1.121994 1 109.500000 1 59.929436 1 2 1 6
H 1.121994 1 109.500000 1 300.070564 1 2 1 6
H 0.991989 1 109.500000 1 179.999999 1 5 4 2
```

ВАРИАНТ 12

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513092D
7 6 0 0 0 0 0 0 0 0999 V2000
  13.5419 -8.3585 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  13.5419 -7.5308 0.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0
  12.8250 -7.1151 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  12.1118 -7.5308 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  14.2632 -6.2915 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  14.2632 -7.1150 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  14.9775 -7.5277 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 3 1 0 0 0 0
3 4 1 0 0 0 0
5 6 2 0 0 0 0
6 2 1 0 0 0 0
2 1 1 0 0 0 0
6 7 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.445999 1 0.000000 0 0.000000 0 1 0 0
C 1.540009 1 109.500000 1 0.000000 0 2 1 0
C 1.540009 1 109.500000 1 60.000000 1 3 2 1
O 3.531514 1 29.907623 1 225.454373 1 1 2 3
C 1.540009 1 109.500000 1 120.070564 1 2 1 3
C 1.540009 1 109.500000 1 299.929436 1 3 2 1
O 1.410004 1 109.500000 1 300.000000 1 6 2 1
H 1.027985 1 109.500000 1 180.000000 1 1 2 3
H 1.027985 1 109.500000 1 59.929436 1 1 2 3
H 1.121994 1 109.500000 1 59.929436 1 2 1 9
H 1.121994 1 109.500000 1 180.070564 1 3 2 1
H 1.121994 1 109.500000 1 179.999999 1 4 3 2
H 1.121994 1 109.500000 1 59.929436 1 4 3 2
```

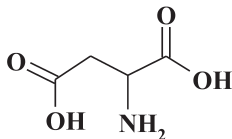
H	1.121994	1	109.500000	1	300.070564	1	4	3	2
H	1.121994	1	109.500000	1	180.000000	1	7	3	2
H	1.121994	1	109.500000	1	59.929436	1	7	3	2
H	1.121994	1	109.500000	1	300.070564	1	7	3	2
H	0.991989	1	109.500000	1	180.000000	1	8	6	2

ВАРИАНТ 13

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513092D
8 7 0 0 0 0 0 0 0 0999 V2000
  -2.6904 -12.42930.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
  -3.3653 -11.96670.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0
  -1.9202 -12.15660.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -2.6613 -13.28640.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -4.0756 -12.36540.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -3.3877 -11.14690.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -1.9202 -11.34200.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  -1.2465 -12.63280.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0 0 0 0
1 3 1 0 0 0 0
1 4 1 0 0 0 0
2 5 1 0 0 0 0
2 6 1 0 0 0 0
3 7 2 0 0 0 0
3 8 1 0 0 0 0
M END
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
N 0.000000 0 0.000000 0 0.000000 0 0 0 0
C 1.469900 1 0.000000 0 0.000000 0 1 0 0
C 1.519900 1 109.470900 1 0.000000 0 2 1 0
O 1.220000 1 119.999900 1 179.999917 1 3 2 1
C 1.540000 1 109.471000 1 240.000500 1 2 1 3
C 1.519900 1 109.471000 1 180.000003 1 5 2 1
O 1.219900 1 120.000000 1 179.999953 1 6 5 2
N 1.320000 1 119.999900 1 359.999953 1 6 5 2
O 1.499900 1 120.000000 1 359.999917 1 3 2 1
H 0.990000 1 119.999900 1 59.999700 1 1 2 3
```

H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

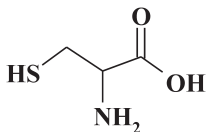
ВАРИАНТ 14

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```
-ISIS- 10130513092D
15 16 0 0 0 0 0 0 0 0999 v2000
  14.9775 -7.5277 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3.4180 -13.00260.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  3.4180 -12.17980.0000 C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0
  2.7082 -11.77040.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  4.1280 -10.96020.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  4.1280 -11.77050.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  2.0851 -12.30430.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  1.2875 -12.11740.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  2.1447 -13.12080.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0.8641 -12.84090.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  1.3876 -13.44400.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0.0314 -12.82740.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 -0.3822 -12.10280.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0.0455 -11.37940.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0.8823 -11.38870.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  4.8423 -12.18320.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 3 1 0 0 0 0
4 5 2 0 0 0 0
5 2 1 0 0 0 0
2 1 1 0 0 0 0
6 7 1 0 0 0 0
6 8 2 0 0 0 0
7 9 2 0 0 0 0
8 10 1 0 0 0 0
9 10 1 0 0 0 0
3 6 1 0 0 0 0
9 11 1 0 0 0 0
11 12 2 0 0 0 0
12 13 1 0 0 0 0
13 14 2 0 0 0 0
14 7 1 0 0 0 0
5 15 1 0 0 0 0
```

M END

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

H	0.960000	1	109.471000	1	180.000000	1	9	3	2
N	0.000000	0	0.000000	0	0.000000	0	0	0	0
C	1.445999	1	0.000000	0	0.000000	0	1	0	0
O	3.100295	1	48.053010	1	0.000000	0	1	2	0
C	1.540009	1	109.500000	1	26.970371	1	2	1	3
O	1.410004	1	109.500000	1	60.000000	1	4	2	1
H	1.027985	1	109.500000	1	180.000000	1	1	2	4
H	1.027985	1	109.500000	1	59.929436	1	1	2	4
H	1.121994	1	109.500000	1	59.929436	1	2	1	6
H	1.121994	1	109.500000	1	300.070564	1	2	1	6
H	0.991989	1	109.500000	1	179.999999	1	5	4	2

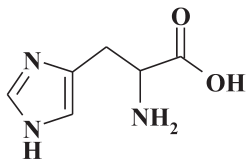
ВАРИАНТ 15

1. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

```

-ISIS- 10130513092D
13 13 0 0 0 0 0 0 0 0999 v2000
  9.0505  -13.28800.0000  N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.0485  -12.43640.0000  C 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0
  8.4196  -11.89830.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.8400  -11.36310.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  9.8254  -12.16040.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  7.7026  -12.31000.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  7.7026  -13.13750.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  6.9940  -13.54930.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  6.2812  -13.13350.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  6.2811  -12.31020.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  6.9938  -11.89840.0000  C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  5.5683  -13.54940.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  10.4950 -12.64240.0000  O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 3 1 0 0 0 0
4 5 2 0 0 0 0
5 2 1 0 0 0 0
2 1 1 0 0 0 0
6 7 1 0 0 0 0
7 8 2 0 0 0 0
8 9 1 0 0 0 0
9 10 2 0 0 0 0
10 11 1 0 0 0 0
11 6 2 0 0 0 0
3 6 1 0 0 0 0
9 12 1 0 0 0 0
5 13 1 0 0 0 0
M END
  
```

2. Пронумеровать молекулу по алгоритму Моргана и записать в коде SMILES, SLN и WLN.



3. Нарисовать молекулу и пронумеровать по алгоритму Моргана, записать в коде SMILES, SLN и WLN.

N	0.000000	0	0.000000	0	0.000000	0	0	0	0
C	1.469900	1	0.000000	0	0.000000	0	1	0	0
C	1.519900	1	109.470900	1	0.000000	0	2	1	0
O	1.220000	1	119.999900	1	179.999917	1	3	2	1
C	1.540000	1	109.471000	1	240.000500	1	2	1	3
C	1.519900	1	109.471000	1	180.000003	1	5	2	1
O	1.219900	1	120.000000	1	179.999953	1	6	5	2
N	1.320000	1	119.999900	1	359.999953	1	6	5	2
O	1.499900	1	120.000000	1	359.999917	1	3	2	1
H	0.990000	1	119.999900	1	59.999700	1	1	2	3
H	1.010000	1	109.470993	1	292.239097	1	1	2	3
H	1.090000	1	109.471400	1	180.000000	1	2	1	10
H	1.090000	1	109.471000	1	299.999403	1	5	2	1
H	1.089900	1	109.471400	1	59.999703	1	5	2	1
H	1.010000	1	120.000000	1	180.000000	1	8	6	5
H	1.010000	1	120.000000	1	359.999900	1	8	6	5
H	0.960000	1	109.471000	1	180.000000	1	9	3	2

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Васильев П. М. Языки фрагментарного кодирования структуры соединений для компьютерного прогноза биологической активности / П. М. Васильев, А. А. Спасов // Рос. хим. журн. — 2006. — Т. 50, № 2. — С. 108–127.

Helson H. E. Structure Diagram Generation / H. E. Helson // Reviews in Computational Chemistry. New York : Wiley-VCH Press, 1999. — Vol. 13. — P. 313–398. — ISBN 978-0-4713-3135-3.

InChI — the worldwide chemical structure identifier standard // InChI Trust : [site]. — URL: <https://www.inchi-trust.org/inchi-post/inchi-the-worldwide-chemical-structure-identifier-standard/> (accessed: 26.04.2019).

SMARTS — A Language for Describing Molecular Patterns // Daylight Headquarters : [site]. — URL: <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> (accessed: 26.04.2019).

SMILES — A Simplified Chemical Language // Daylight Headquarters : [site]. — URL: <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> (accessed: 26.04.2019).

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules / D. Weininger // J. Chem. Inf. Comput. Sci. — 1988. — Vol. 28, № 1. — С. 31–36.

Weininger D. SMILES. 2. Algorithm for generation of unique SMILES notation / D. Weininger, A. Weininger, J. L. Weininger // J. Chem. Inf. Comput. Sci. — 1989. — Vol. 29, № 2. — P. 97–101.

Учебное издание

Нейн Юлия Ивановна
Иванцова Мария Николаевна

КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКОЙ ИНФОРМАЦИИ

Учебное пособие

Заведующий редакцией	<i>М. А. Овечкина</i>
Редактор	<i>Н. В. Чапаева</i>
Корректор	<i>Н. В. Чапаева</i>
Оригинал-макет	<i>Л. А. Хухаревой</i>

Подписано в печать 29.05.2020. Формат 60 × 84^{1/16}.
Бумага офсетная. Цифровая печать. Усл. печ. л. 8,37.
Уч.-изд. л. 6,2. Тираж 30 экз. Заказ 78

Издательство Уральского университета
Редакционно-издательский отдел ИПЦ УрФУ
620083, Екатеринбург, ул. Тургенева, 4
Тел.: +7 (343) 389-94-79, 350-43-28
E-mail: rio.marina.ovechkina@mail.ru

Отпечатано в Издательско-полиграфическом центре УрФУ
620083, Екатеринбург, ул. Тургенева, 4
Тел.: +7 (343) 358-93-06, 350-58-20, 350-90-13
Факс: +7 (343) 358-93-06
<http://print.urfu.ru>

ДЛЯ ЗАМЕТОК

ДЛЯ ЗАМЕТОК

