Научная визуализация, 2018, том 10, номер 3, страницы 133 - 143, DOI: 10.26583/sv.10.3.10

Комплекс программ визуализации молекулярных структурных графов органических соединений

В.В. Бондарь[1,A], Е.Г. Винокуров[2,A,B], В.П. Мешалкин[3,B,C], Л.А. Григорян[4,A]

[A] Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)
[B] Российский химико-технологический университет имени Д. И. Менделеева (РХТУ им. Д. И. Менделеева)
[C] Институт общей и неорганической химии им. Н. С. Курнакова РАН

[1] ORCID: 0000-0002-6708-9025, Levgr2@yandex.ru
[2] ORCID: 0000-0002-5376-0586
[3] ORCID: 0000-0001-6956-6705
[4] ORCID: 0000-0002-6535-6185

Аннотация

Представлен комплекс программ визуализации структурных графов органических соединений, генерируемых по запросу пользователя. Рассмотрена архитектура программного комплекса. Дано описание процедуры лингвистического анализа химико-структурных данных с использованием классификации морфем химической номенклатуры и контекстных правил укорачивающей грамматики. Отмечено прикладное значение программного комплекса. Приведены примеры визуализации ряда структурных графов из различных разделов номенклатуры органических соединений ИЮПАК.

Ключевые слова: структурный граф, химическая номенклатура, анализ данных, научная визуализация, прикладная программа.

1. Введение

Комплекс программ «Номенклатурный Генератор» предназначен для автоматической генерации молекулярных структурных графов (МСГ) органических соединений по вводимому пользователем систематическому наименованию [1–3]. «Номенклатурный Генератор» позволяет генерировать МСГ для следующих соединений по номенклатуре органических соединений ИЮПАК: алифатические соединения; моноциклические соединения; соединения с функциональными группами, называемые по заместительной номенклатуре; соединения, называемые по заменительной или «а»-номенклатуре; соединения, называемые по расширенной системе Ганча-Видмана; отдельные подклассы гетероциклических и ароматических соединений; некоторые соединения, сохраняющие тривиальные наименования.

Автоматизированная генерация МСГ по названиям органических соединений упрощает труд научных работников, занимающихся пополнением баз данных (БД) молекулярных структур химических и фармакологических веществ. Визуализация генерируемых МСГ повысит эффективность указанных БД, сделав содержащийся в них контент более наглядным, полным и удобным для восприятия и дальнейшего использования. Комплекс программ «Номенклатурный Генератор» может быть также полезен в учебном процессе по химии в качестве виртуального тренажера для студентов химико-технологического и инженерного профиля и учащихся старших классов общеобразовательных

школ в свете новых стандартов ФГОС [2–4].

2. Комплекс программ «Номенклатурный Генератор»

Ключевым преимуществом комплекса программ «Номенклатурный Генератор» по сравнению с зарубежными аналогами является возможность обработки запросов на русском естественном языке [3, 5–6]. К числу аналогов можно отнести, прежде всего, программный пакет ChemOffice разработчика CambridgeSoft [7], пакет MDL (Beilstein) AutoNom [8–9], программу ACD/Name компании ACD/Labs [10], модуль Lexichem компании OpenEye Scientific Software [11] и др. [12], однако эти системы, за исключением модуля Lexichem, не предусматривают возможности работы с русскоязычной версией химической номенклатуры, а доступ к Lexichem существенно ограничен условиями лицензионного соглашения. В связи с этим, важно отметить, что проект «Номенклатурный Генератор» является некоммерческой разработкой с открытым программным кодом.

Принципы смысловой обработки текстов, задействованные в алгоритмах комплекса программ «Номенклатурный Генератор», были заложены в работах В. В. Кафарова [13], Г. Стефанопулоса [14], Н. С. Зефирова [15–17], В. А. Коптюга [18] и др. [19–21]. Основное отличие «Номенклатурного Генератора» от работ предшественников заключается в использовании алгоритмом программного комплекса укорачивающей грамматики GRM (грамматики свёртки), состоящей из нескольких групп контекстных правил, последовательное обращение к которым преобразует систематическое название химического соединения в молекулярно-структурный граф.

Комплекс программ «Номенклатурный Генератор» имеет трёхмодульную функциональную архитектуру (см. рис. 1), включающую основной модуль-номенклатор, который непосредственно генерирует МСГ по названию органического соединения. Вспомогательными модулями являются: модуль-переводчик, который осуществляет перевод названий органических соединений с русского языка на английский и с английского на русский [3], и модуль-конкорданс, выполняющий функцию массива-накопителя МСГ.

Рис. 1. Обобщенная блок-схема функциональной архитектуры комплекса программ «Номенклатурный Генератор»

Взаимодействие между модулями осуществляет морфо-синтаксический алгоритм, разработанный с учетом правил химической номенклатуры ИЮПАК [22].

Блок-схема информационно-программного обеспечения комплекса программ «Номенклатурный Генератор» представлена на рис. 2.

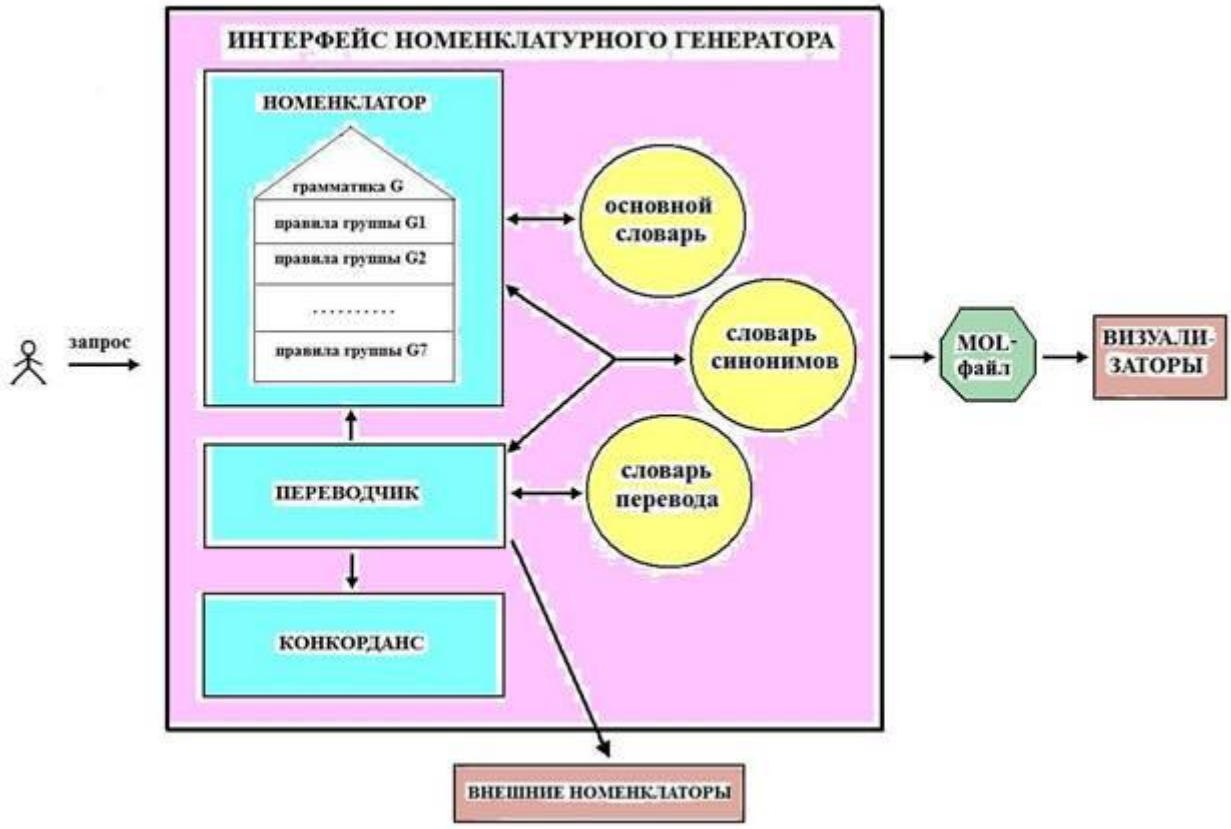


Рис. 2. Блок-схема информационно-программного обеспечения комплекса программ «Номенклатурный Генератор»

Алгоритм комплекса программ «Номенклатурный Генератор» включает следующие основные этапы:

1) ввод пользователем номенклатурного названия органического соединения;

2) автоматический анализ введённого названия с использованием процедуры разбиения его на составные части (морфемы и субморфы) [2];

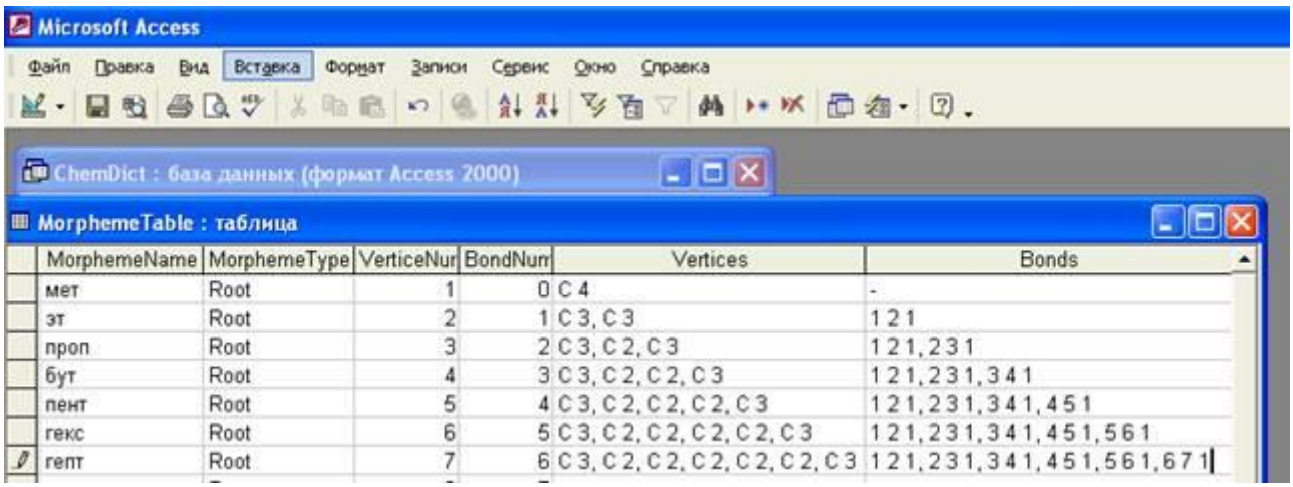
3) выявление внутренней семантико-синтаксической структуры введённого названия;

4) обращение к встроенным специализированным словарям химических морфем [23];

5) автоматическая генерация МСГ органического соединения, соответствующего введённому названию, в стандартизованном формате mol-файла [1–2, 24];

6) взаимодействие с внешними универсальными визуализаторами (не входящими в оболочку программного комплекса) для отображения полученного МСГ.

Алгоритм реализован на языке программирования C++. При функционировании комплекса программ «Номенклатурный Генератор» используются разработанные нами специализированные словари химических морфем, в том числе химико-лингвистический словарь, словарь перевода и синонимический словарь [25]. Словари программно реализованы в редакторе баз данных Microsoft Access и интегрированы в оболочку комплекса программ в качестве подключаемых библиотек. Пример экранной формы таблицы химико-лингвистического словаря представлен на рис. 3:



MorphemeName	MorphemeType	VerticeNur	BondNum	Vertices	Bonds
мет	Root	1	0	C 4	-
эт	Root	2	1	C 3, C 3	1 2 1
проп	Root	3	2	C 3, C 2, C 3	1 2 1, 2 3 1
бут	Root	4	3	C 3, C 2, C 2, C 3	1 2 1, 2 3 1, 3 4 1
пент	Root	5	4	C 3, C 2, C 2, C 2, C 3	1 2 1, 2 3 1, 3 4 1, 4 5 1
гекс	Root	6	5	C 3, C 2, C 2, C 2, C 2, C 3	1 2 1, 2 3 1, 3 4 1, 4 5 1, 5 6 1
гепт	Root	7	6	C 3, C 2, C 2, C 2, C 2, C 2, C 3	1 2 1, 2 3 1, 3 4 1, 4 5 1, 5 6 1, 6 7 1

Рис. 3. Фрагмент таблицы химико-лингвистического словаря в редакторе баз данных Microsoft Access

Основным этапом алгоритма «Номенклатурного Генератора» является процедура морфемного анализа с последующей свёрткой выявленных компонентов названия органического соединения в пустое слово и одновременным построением единого структурного графа. Алгоритм опирается на лингвистический аппарат, адаптированный к специфике поставленной задачи. Ключевыми элементами лингвистического аппарата являются:

- усовершенствованная нами методика классификации химических морфем и субморфов, восходящая к работам А. М. Цукермана – Г. Э. Влэдуца – Г. Г. Стецюры [26–29], М. М. Ланглебен [30–32] и др., обзор которых дан в нашей работе [1];
- укорачивающая грамматика GRM, представляющая собой упорядоченную совокупность т.н. контекстных правил [2].

Отличие предложенной нами методики классификации химических морфем (см. табл. 1) от предшествующих вариантов состоит, прежде всего, в разбиении всех типов химических морфем на два класса – словарных (семантико-синтаксический класс) и служебных (морфо-синтаксический класс). Типология морфем внутри классов определяется их ролью при составлении названия химических соединений. Так, морфемы, отнесённые к корневому типу, кодируют фрагменты структурных цепочек химического соединения, морфемы суффиксального типа отвечают за модификации структурного графа, касающиеся кратности связи между атомами, наличия в соединении функциональных характеристических групп и др. [1]

Табл. 1. Классификация химических морфем

Синтаксический тип	**Класс**	**Семантическая роль** (химико-номенклатурный смысл)
Root (Корень)	Словарный (CC[1])	Структурная цепочка или ее фрагмент
Suffix (Суффикс)	Словарный (CC)	Кратность связи; главная функциональная группа; модификации структурной цепочки или связей
Prefix (Префикс)	Словарный (CC)	Функциональная группа; модификации структурной цепочки и др.
Multi (Количественная приставка)	Словарный (CC)	Кратность повторяющихся фрагментов структурной цепочки, модификаторов связи и др.

Hetero (Гетеро-префикс)	Словарный (СС)	Гетероатом внутри структурной цепочки
Hydro (Префикс гидрирования)	Словарный (СС)	Гидрирование атомов структурной цепочки
PeriodicSymbol (Символ периодической системы)	Словарный (СС)	Атом в производных соединениях неорганической химии и природных соединениях
Unspec (Неспецифический компонент)	Словарный (СС)	Переводной компонент названия с неуточненной химико-структурной функцией
Locant (Локант)	Словарный (СС)	Указатель модифицируемой вершины структурного графа, стыковой вершины, грани и др.
Comma (Запятая)	Служебный (МС[2])	Разделитель между локантами внутри комплекса, между основной и инвертированной частью названия и др.
Hyphen (Дефис)	Служебный (МС)	Разделитель, обособляющий комплексы локантов, слабостыкующиеся морфемы и др., показатель отрицательного заряда
Point (Точка)	Служебный (МС)	Разделитель между локантами в мостиковых соединениях и др.
OpenBracket (Открывающая скобка)	Служебный (МС)	Показатель левой границы сложного комплекса морфем в иерархических структурах
CloseBracket (Закрывающая скобка)	Служебный (МС)	Показатель правой границы сложного комплекса морфем в иерархических структурах
Asterisk (Астериск)	Служебный (МС)	Модификатор стереохимического префикса
Apostrophe (Штрих)	Служебный (МС)	Модификатор локанта
Colon (Двоеточие)	Служебный (МС)	Разделитель сложных групп локантов, показатель соотношения мер и др.
Plus (Плюс)	Служебный (МС)	Показатель положительного заряда

[1] СС – семантико-синтаксический класс морфем.
[2] МС – морфо-синтаксический класс морфем.

Грамматика свёртки GRM насчитывает 7 групп контекстных правил, каждая из которых предназначена для обработки определенных комбинаций химических морфем внутри названия химического соединения. Циклическое применение контекстных правил приводит к поэтапной свёртке названия химического соединения в пустое слово, сопровождающейся восстановлением соответствующей ему химической структуры: 1) выявляются присутствующие в соединении структурные цепочки; 2) цепочки модифицируются в зависимости от наличия гетероатомов, кратных связей, функциональных характеристических групп; 3) осуществляется соединение модифицированных цепочек в единый МСГ всего соединения согласно расположению стыковых атомов в цепочках. [33]

В качестве примера рассмотрим анализ названия химического соединения «*3-Метилпентан*». Это название состоит из морфем «*3*», «-», «*мет*», «*ил*», «*пент*», «*ан*». На начальной стадии алгоритм преобразует данное название в последовательность вида {$Locant_1$(3), Hyphen, $Root_1$(мет), $Suffix_1$(ил), $Root_2$(пент), $Suffix_2$(ан)}. После считывания информации о химических морфемах из специализированного словаря, последовательность примет вид {$Locant_1$(3), Hyphen, $Root_1$(CH_4), $Suffix_1$(*свободная_одинарная_связь*), $Root_2$(CH_3 - CH_2 - CH_2 - CH_2 - CH_3), $Suffix_2$(*одинарная_связь*)}. Далее задействуется грамматика свертки GRM.

Применение контекстного правила 2R2 придаст последовательности вид {Locant$_1$(3), Hyphen, Root'$_1$(CH_3-), Root$_2$(CH_3 - CH_2 - CH_2 - CH_2 - CH_3), Suffix$_2$(*одинарная_связь*)}. Повторное применение правила 2R2 даст результат {Locant$_1$(3), Hyphen, Root'$_1$(CH_3-), Root'$_2$(CH_3 - CH_2 - CH_2 - CH_2 - CH_3)}. Финальное применение правила 7R2 свернёт последовательность в единый МСГ:

{Root''$_2$(CH_3 - CH_2 - CH_2 - CH_2 - CH_3)
|
CH_3 }.

Результат будет выведен в типизированный файл в формате MOL, содержащий информацию о вершинах и рёбрах МСГ.

Комплекс программ «Номенклатурный Генератор» совместим с различными версиями операционной системы Windows, начиная от Windows XP и заканчивая Windows 10. Сетевая версия «Номенклатурного Генератора», допускающая ввод данных через интернет, находится в стадии разработки.

Экранная форма, изображающая интерфейс комплекса программ, представлена на рис. 4:

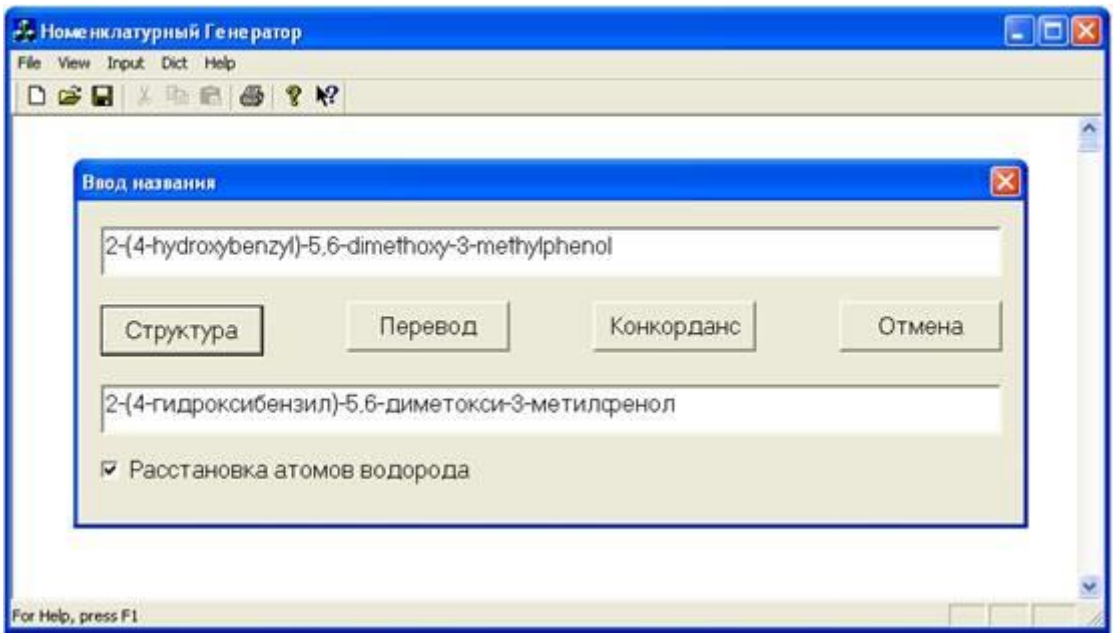


Рис. 4. Интерфейс комплекса программ «Номенклатурный Генератор»

Практическая значимость и актуальность проекта подтверждены актом о внедрении (ЗАО ЦИНТЭНСИ) и свидетельством Роспатента о государственной регистрации комплекса программ «Номенклатурный Генератор» [24].

3. Примеры визуализации

В качестве примера рассмотрим результат обработки «Номенклатурным Генератором» ряда названий органических соединений из вышеперечисленных разделов номенклатуры ИЮПАК.

Так, для названия органического соединения *5-Метил-4-пропилнонан*, относящегося к классу алифатических соединений, «Номенклатурный Генератор» построит МСГ, показанный на рис. 5:

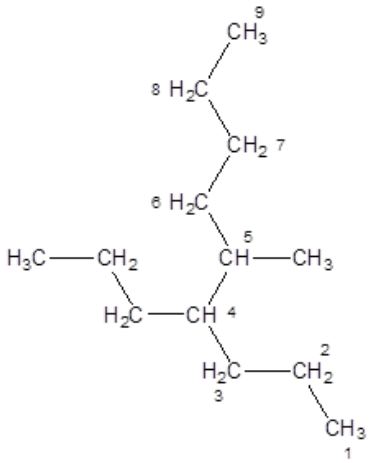


Рис. 5. МСГ соединения 5-Метил-4-пропилнонан

По названию соединения *2,8-Диметил-1,3,4-трипропилциклоундекан* (класс моноциклических соединений) будет сгенерирован МСГ, показанный на рис. 6:

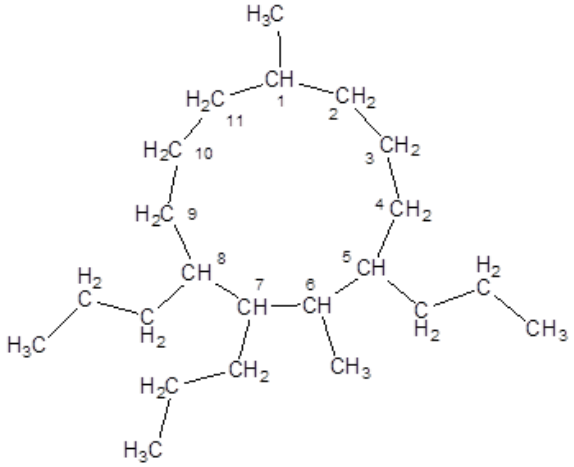


Рис. 6. МСГ соединения 2,8-Диметил-1,3,4-трипропилциклоундекан

По названию соединения *6-(5-Гидроксипент-3-ен-1-инил)ундека-2,4,7-триен-9-ин-1,11-диол* (заместительная номенклатура) комплекс программ сгенерирует МСГ, показанный на рис. 7:

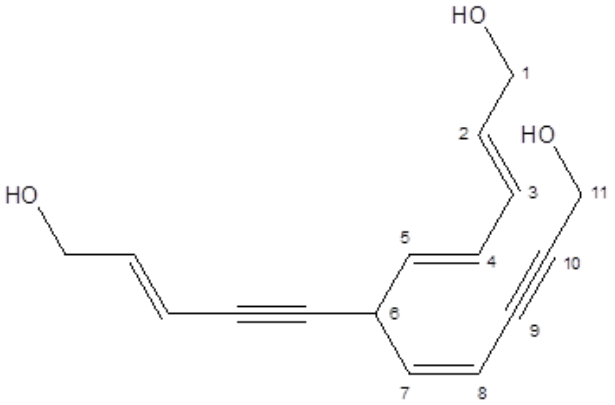


Рис. 7. МСГ соединения 6-(5-Гидроксипент-3-ен-1-инил)ундека-2,4,7-триен-9-ин-1,11-диол

По названию соединения *6,15-Диокса-3,11-дисиладокозан* (заменительная номенклатура) будет построен МСГ, показанный на рис. 8:

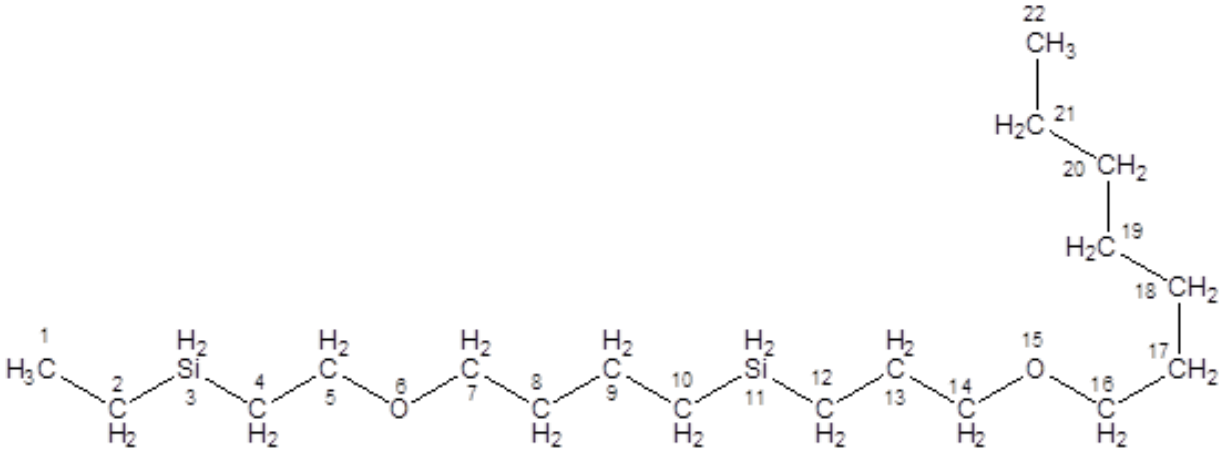


Рис. 8. МСГ соединения 6,15-Диокса-3,11-дисиладокозан

По названию соединения *2,3-Дигидро-6H-1,5,3-диазасилонин* (расширенная система Ганча–Видмана) будет построен МСГ, показанный на рис. 9:

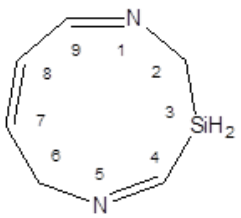


Рис. 9. МСГ соединения 2,3-Дигидро-6H-1,5,3-диазасилонин

По названию соединения *2,3,4,7,8,9-Гексаметилгептален* (класс ароматических соединений) будет сгенерирован МСГ, показанный на рис. 10:

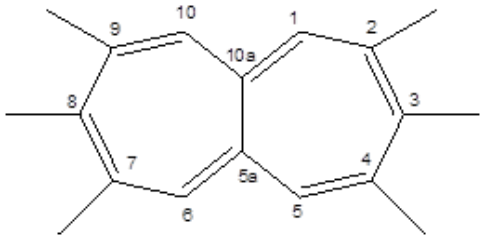


Рис. 10. МСГ соединения 2,3,4,7,8,9-Гексаметилгептален

По названию соединения *Ацетон* (тривиальное наименование) будет построен МСГ, показанный на рис. 11:

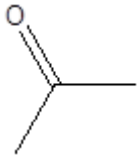


Рис. 11. МСГ соединения Ацетон

Приведенные МСГ могут быть отображены, например, посредством внешнего визуализатора HyperChem версии 5.02 и выше. Комплекс программ «Номенклатурный Генератор» совместим и с другими внешними визуализаторами, в т.ч. Isis/Draw. Выбор оптимального режима визуализации оставлен на усмотрение пользователя.

4. Заключение

Комплекс программ «Номенклатурный Генератор», сочетающий элементы лингвистического анализа данных и структурного синтеза молекулярных графов, способен обрабатывать до нескольких миллионов систематических названий органических соединений, вводимых пользователем на русском естественном языке, с последующей визуализацией полученных результатов.

Благодарность

Исследование выполнено при поддержке Министерства образования и науки РФ в рамках выполнения базовой части государственного задания.

Список литературы

1. Bondar’ V. V., Vinokurov E. G., Grigoryan L. A. Shortening Grammar Based on the Renewed Classification of Chemical Nomenclature Morphemes for Use in the «Nomenclature Generator» Software Complex // Automatic Documentation and Mathematical Linguistics, 2014, Vol. 48, No. 4, pp. 199–207.

2. Мешалкин В. П., Бондарь В. В., Винокуров Е. Г., Григорян Л. А. Морфосинтаксический алгоритм компьютерного анализа систематических названий базовых и модифицированных алифатических соединений // Теоретические основы химической технологии, 2017, т. 51, № 5.

3. Бондарь В. В., Григорян Л. А., Немировская И. Б. Программа перевода систематических названий химических соединений с русского на английский язык. // Хим. технол. 2007, N 2, с. 93–96.

4. Осипов Г. В., Стриханов М. Н., Шереги Ф. Э. Компетентностное образование инженеров-инноваторов // Социология образования. 2015. № 4. С. 4–27.

5. Sayle R. Foreign language translation of chemical nomenclature by computer. – Journal of Chemical Information and Modeling, 2009, v. 49, № 3, pp. 519–530.

6. Sayle R., Xie P. H., Muresan S. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. – Journal of Chemical Information and Modeling, 2012, v. 52, № 1, pp. 51–62.

7. Zielesny A. Chemistry software package: ChemOffice Ultra 2005. – Journal of Chemical Information and Modeling, 2005, v. 45, № 5, pp. 1474–1477.

8. Wiśniewski J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names: 1. General Design. – Journal of Chemical Information and Computer Sciences, 1990, v. 30, № 3, pp. 324–332.

9. Goebels L., Lawson A. J., Wiśniewski J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names: 2. Nomenclature of Chains and Rings. – Journal of Chemical Information and Computer Sciences, 1991, v. 31, № 2, pp. 216–225.

10. Williams A., Yerin A. The Need for Systematic Naming Software Tools for Exchange of Chemical Information. – Molecules, 1999, v. 4, pp. 255–263.

11. Sayle R. Foreign language translation of chemical nomenclature by computer. – Journal of Chemical Information and Modeling, 2009, v. 49, № 3, pp. 519–530.

12. Brecher J. Name=Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature. – Journal of Chemical Information and Computer Sciences, 1999, v. 39, № 6, pp. 943–950.

13. Кафаров В. В., Мешалкин В. П., Дюкова Е. А. Принципы разработки семантико-математических моделей понимания смысла технологических текстов // Докл. АН СССР. 1989. Т. 306. № 4. С. 916.

14. Stephanopoulos G., Han C. Advances in Chemical Engineering: Intelligent Systems // Process Engineering Paradigms from Design and Operations. Academic Press, 1995. V. 21–22. P. 297.

15. Баскин И. И., Палюлин В. А., Зефиров Н. С. Применение искусственных нейронных сетей для прогнозирования свойств химических соединений // Нейрокомпьютеры: разработка, применение. 2005. № 1-2. С. 98.

16. Melnikov A. A., Palyulin V. A., Zefirov N. S. Generation of molecular graphs for QSAR studies: an approach based on supergraphs. – Journal of Chemical Information and Modeling, 2007, v. 47, № 6, pp. 2077–2088.

17. Tetko I. V., Tanchuk V. Yu., Prokopenko V. V., Gasteiger J., Todeschini R., Mauri A., Livingstone D., Ertl P., Palyulin V. A., Radchenko E. V., Zefirov N. S., Makarenko A. S. Virtual computational chemistry laboratory – design and description // Journal of Computer-Aided Molecular Design. 2005. Т. 19. № 6. С. 453-463.

18. Дерендяев Б. Г., Коптюг В. А., Лебедев К. С., Шарапова О. Н. Машинно-информационная система на базе каталогов полных масс-спектров // Автометрия. 1979. № 4. С. 3.

19. Xu R., Yang Y. Cross-lingual Distillation for Text Classification // arXiv preprint arXiv:1705.02073, 2017.

20. Lowe D. M., Sayle R. A. LeadMine: A grammar and dictionary driven approach to entity recognition // Journal of Cheminformatics, 2015, v. 7, S5.

21. Мирошников А. Н., Осипов А. Л. Программные средства представления химических структур // Сб. «Научно-технический прогресс: актуальные и перспективные направления будущего»: сборник материалов II Международной научно-практической конференции: в 2 томах. 2016. С. 108–110.

22. International Union of Pure and Applied Chemistry. URL: https://iupac.org/what-we-do/books/color-books (проверено 30 июня 2017 г.).

23. Григорян Л. А. Разработка словарей морфем химической номенклатуры // Вестник РГГУ, № 8 (130), серия «Филологические науки. Языкознание» / Московский лингвистический журнал, т. 16, 2014. – С. 139–149.

24. Григорян Л. А., Винокуров Е. Г., Бондарь В. В., Марголин Л. Н., Фарафонов В. В., Королева Л. М. Программный комплекс «Номенклатурный Генератор», предназначенный для преобразования названий органических соединений в MOL-формат, отражающий структуру молекулярного графа. Свидетельство о государственной регистрации программы для ЭВМ № 2014619365. Правообладатель ФГБУН ВИНИТИ РАН (RU); заявка № 2014617378; дата поступления 25.07.2014; дата гос. регистрации в Реестре программ для ЭВМ 15.09.2014.

25. Григорян Л. А. Разработка словарей морфем химической номенклатуры // Вестник РГГУ, № 8 (130), серия «Филологические науки. Языкознание» / Московский лингвистический журнал, т. 16, 2014. – С. 139–149.

26. Влэдуц Г. Э. Некоторые вопросы научной информации в области химии. – М.: Институт научной информации АН СССР, 1958. – 134 с.

27. Цукерман А. М., Стецюра Г. Г. Об автоматизации перевода названия химических органических соединений в стандартную форму и структурных формул в систематические наименования // Сообщ. лаборатории электромоделирования. Вып. 1. – М.: Институт научной информации АН СССР, 1960, с. 241.

28. Стецюра Г. Г., Цукерман А. М. Автоматический перевод названия химических органических соединений в формулы // НТИ, 1962, № 3, с. 17–19.

29. Цукерман А. М. Номенклатура органических соединений и номенклатурный перевод. – М., 1966. – 253 с.

30. Ланглебен М. М. К лингвистическому описанию номенклатуры органической химии // НТИ, 1967, № 1, с. 13–22.

31. Ланглебен М. М. Структура номинативных сочетаний в специальном фрагменте русского химического языка : дис. канд. филол. наук / М. М. Ланглебен . – М.: ВИНИТИ, 1970. – 257 с.

32. Ланглебен М. М. Опыт построения метаязыка для описания квазилингвистической семиотической системы // Сб. «Исследования по математической лингвистике, математической логике и информационным языкам» / под ред.: Д. А. Бочвар, Ю. А. Шрейдер ; АН СССР. – М.: Наука, 1972. – с. 96–146.

33. Meshalkin V. P., Bondar' V. V., Vinokurov E. G., Grigoryan L. A. Morphosyntactic Algorithm for Computer-Assisted Analysis of Systematic Names of Fundamental and Modified Aliphatic Compounds // Theoretical Foundations of Chemical Engineering, 2017, Vol. 51, No. 5, pp. 752–758.