13.

АВТОМАТИЧЕСКОЕ ПОРОЖДЕНИЕ СТРУКТУРЫ ПО НАЗВАНИЮ ХИМИЧЕСКОГО СОЕДИНЕНИЯ

"STRUCTURE-BY-NAME" AUTOMATICAL GENERATION FOR THE CHEMICAL COMPOUNDS

Л.А. Григорян *(Levgr2@yandex.ru)*

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Разработана новая версия программы "Номенклатурный Анализатор", переводящей вводимые пользователем названия химических соединений, данные в систематической номенклатуре ИЮПАК, в молекулярные графы. Алгоритм программы основан на морфемном членении названий химических соединений на химически-осмысленные составные части-морфемы.

Введение

Данная программа является развитием проекта проф. В.К.Финна, основу которого заложила в 1999 г. лингвист Е.А.Уткина [1].

Суть алгоритма сводится к морфемному дроблению вводимого пользователем названия химического соединения на составные части-морфемы, обладающие определённой химической семантикой. Взаимное расположение этих морфем задаёт структуру соответствующего названию молекулярного графа. Фактически данная проблема представляет собой лингвистический анализ сложных слов, включающий морфологический и синтаксический этапы.

Предыстория вопроса

Из многочисленных химических номенклатур [2] наиболее известны номенклатура Международного союза теоретической и прикладной химии (International Union of Pure and Applied Chemistry, ИЮПАК) [3] и номенклатура Американского химического общества (American Chemical Society) - номенклатура CAS [4]. Русскоязычным аналогом номенклатуры ИЮПАК является номенклатура Всероссийского института научной и технической информации (ВИНИТИ РАН), на основе которой, собственно, и создаётся Анализатор.

Углубление знаний о строении химических соединений требует отражения новых сведений в названиях, что стимулирует совершенствование химических номенклатур. В них вносятся дополнения, изменения и уточнения. Общая тенденция в этом процессе - стремление к единообразию, стандартизации и систематизации названий химических соединений.

Вообще смысл химической номенклатуры заключается в том, чтобы для каждого химического соединения по его названию могла бы быть восстановлена его химическая структура, а по химической структуре, в свою очередь, могло бы быть построено официально принятое название, максимально точно эту структуру отражающее. Следовательно, принятые в номенклатуре названия соединений должны члениться на отдельные компоненты, наделённые определённой химической информацией. Имея полный набор таких компонентов и руководствуясь принятыми в номенклатуре синтаксическими правилами обращения с ними, можно выстроить из них, как из кубиков, любое верное химическое название.

Что касается "Номенклатурного Анализатора", то эта программа призвана решать обратную задачу. Вводимое пользователем номенклатурное название химического соединения подвергается автоматическому дроблению на составные части (морфемы). Затем, полученная последовательность морфем сопоставляется с заложенными в программу синтаксическими правилами химической номенклатуры. Каждому такому правилу соответствует особый тип восстановления молекулярной структуры данного соединения.

Отдельную проблему составляет ситуация с так называемыми "тривиальными" названиями, то есть с названиями, сложившимися исторически или привычно используемыми в обиходе. Естественно, разложить их на химически осмысленные компоненты или восстановить по ним структурную формулу соединения невозможно. Некоторые химические номенклатуры, как, например, номенклатура ИЮПАК, используют тривиальные названия параллельно с систематическими практически без ограничений. Другие, как номенклатура CAS (создаваемая с учётом требований автоматизированных поисковых систем), максимально исключают тривиальные названия.

Проблема построения систематической химической номенклатуры с чётко определёнными правилами встала перед научным сообществом достаточно давно, ещё в XVIII веке, но особую актуальность приобрела в

середине XIX века, когда были сформулированы основные принципы органической химии, согласно которым важен не только состав атомов в веществе, но также их взаимное расположение и количество связей между ними, то есть важна сама структура вещества.

Для однозначного описания структуры органических веществ необходимо было создать полноценную семиотическую систему с развитыми синтаксисом и семантикой. Одной из первых появилась Женевская номенклатура (1892 г.). Её основным недостатком была неполнота, кроме того, она применялась лишь к достаточно простым соединениям. В 1930 г. Международный союз химии принял Льежскую номенклатуру. По её правилам для одного соединения часто допускалось несколько названий, что приводило к путанице в терминологии.

В Советском Союзе попытку создания современной химической номенклатуры предпринял А.П.Терентьев [5], но его система не была принята, так как требовала отказа от устоявшихся тривиальных названий. В настоящее время, как уже упоминалось, наиболее ходовыми являются номенклатуры IUPAC и CAS. Номенклатуры эти англоязычны, но у них существуют аналоги в различных странах, в том числе и в России. Правила перевода номенклатуры на другие языки просты, потому что фрагменты (морфемы), из которых строятся названия соединений, интернациональны. Теоретическую основу под взаимодействие химической номенклатуры как семиотической системы, выстроенной на искусственном подъязыке, и естественного языка, в который эта система встраивается, заложили работы М.М.Ланглебен [6-9]. Ею же создана грамматика в виде порождающей модели синтаксиса номенклатуры органических соединений. Алгоритм номенклатурного перевода, использованный в "Номенклатурном Анализаторе" Е.А.Уткиной, разработан А.М.Цукерманом [10].

И номенклатура ИЮПАК, и номенклатура CAS - англоязычны, но у них существуют аналоги в различных странах, в том числе и в России. Правила перевода номенклатуры на другие языки просты, потому что фрагменты (морфемы), из которых строятся названия соединений, интернациональны.

Предпосылки к разработке программы "Номенклатурный Анализатор"

С возникновением вездесущих компьютерных систем оказалось возможным значительно упростить процессы, связанные с обработкой больших объёмов информации. То, что раньше делалось вручную, отнимая порой очень много времени, современный компьютер в состоянии осуществить за доли секунды (естественно, если задача алгоритмизована). Всякая систематическая химическая номенклатура содержит конечный набор правил построения названий химических соединений по их структуре и порождения структуры соединения по названию. Совокупность этих правил представляет собой, в сущности, алгоритм, заданный на естественном языке. Перевод этого алгоритма на "язык" компьютера, позволит значительно облегчить труд учёных и переводчиков, имеющих дело со сложными, "многоэтажными" химическими наименованиями, распознать структуру которых с первого взгляда крайне сложно. Человеку больше не придётся держать в голове весь спектр номенклатурных правил с учётом приоритета их применения, а это сведёт к минимуму ошибки при построении сложных названий или структур.

Для большинства химических соединений между систематическим названием и структурой существует взаимное соответствие. Поэтому многие базы данных, имеющие отдельные поля для названия и для структуры содержат дублированную информацию, что отрицательно сказывается на объёме этих баз данных. Возможность в любой момент по структуре соединения получить его название (или наоборот) позволит упростить и сократить такие базы данных.

Сегодня уже имеются компьютерные системы для работы с химическими номенклатурами (например, немецкий пакет AutoNom, канадский номенклатор ACD/Labs, или кембриджский пакет ChemOffice). Но они недёшевы и, кроме того, не удобны для российских пользователей, так как не рассчитаны на русифицированные варианты номенклатур. Потому возникла потребность в создании аналогичной русскоязычной разработки. Е.А.Уткина [1] заложила основы решения первой части задачи - синтеза структуры химического соединения по его названию.

Программа "Номенклатурный Анализатор" представляет собой полноценное программное приложение, способное обрабатывать основные классы органических соединений, и, главное, в отличие от остальных систем такого рода, открытое для доработок.

Типы морфем и лингвистические аспекты задачи

Для успешной реализации поставленной перед "Номенклатурным Анализатором" задачи потребовался морфо-синтаксический подход к предметной области.

Согласно идеям А.М.Цукермана [10] и М.М.Ланглебен [6-9], химическая номенклатура рассматривается как своего рода искусственный язык. В качестве элементарных и неделимых частиц этого языка выступают химические морфемы, служебные знаки и локанты. Всего насчитывается несколько тысяч различных морфем, включая тривиальные (т.е. несистематические, хотя и признаваемые в рамках данной номенклатуры). Служебных знаков существует гораздо меньше - это, прежде всего, дефис, запятая, точка, апостроф, круглые, квадратные и угловые скобки. Локантами называют числа, а также некоторые латинские и греческие символы, исполняющие функцию адресных указателей на тот или иной фрагмент структуры химического соединения (т.е. они представляют собой индексы операций, которые необходимо проделать над морфемами-компонентами химической структуры). С алгоритмической точки зрения удобно рассматривать служебные знаки и локанты как

особые служебные разновидности морфем, хотя, если придерживаться строго лингвистической модели, то правильнее было бы счесть их знаками пунктуации и метатекстовыми единицами.

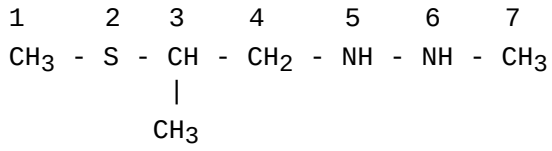
Все морфемы поделены на несколько типов, в соответствии с их различной ролью при построении из них номенклатурного названия.

В основе названия химического соединения чаще всего лежит т.н. "корневая" морфема (в кодировке Анализатора ей присвоено обозначение **Root**). Морфема типа Root описывает базисный углеродный каркас химического вещества - т.е. последовательность атомов углерода со связями между ними.

Если рассматриваемое вещество отличается от базисного, например, кратностью связей или особыми присоединёнными элементами, то к корневой морфеме слева или справа добавляются функциональные морфемы типа **Prefix** или **Suffix**. Они, вместе с прилегающими к ним локантами, указывают, что именно надо изменить в базисной структуре, чтобы построить правильный молекулярный граф.

Кроме того, имеется класс морфем **Multi**, включающий в себя все умножающие приставки. Также существует несколько других, менее употребительных классов морфем.

Рассмотрим, например, название вещества "*3-метил-2-тиа-5,6-диазагептан*", соответствующее следующему структурному графу:



Алгоритм Анализатора разобьёт это название на следующую последовательность морфем: "*3*", "-", "*мет*", "*ил*", "-", "*2*", "-", "*тиа*", "-", "*5*", ",", "*6*", "-", "*ди*", "*аза*", "*гепт*", "ан".

В представлении Анализатора эти морфемы принадлежат следующим типам:

3 **Locant**
- **Hyphen**
мет **Root**
ил **Suffix**
- **Hyphen**
2 **Locant**
- **Hyphen**
тиа **Hetero**
- **Hyphen**
5 **Locant**
, **Comma**
6 **Locant**
- **Hyphen**
ди **Multi**
аза **Hetero**
гепт **Root**
ан **Suffix**

Далее полученная последовательность типов морфем соотносится с имеющимися синтаксическими правилами сочетаемости, заложенными в алгоритм программы, и постепенно преобразуется в сторону упрощения, согласно порождающей грамматике М.М.Ланглебен.

Таким образом по морфеме "*гепт*" восстанавливается базисная структура соединения - углеродная цепь из 7 атомов. Морфема "*ан*" указывает, что все связи в данном соединении - одинарные. Фрагмент "*2-тиа-5,6-диаза*" (называемый *составным гетеропрефиксом*) свидетельствует о том, что в исходном углеродном каркасе вместо второго атома цепочки должен стоять атом серы **S** (ему соответствует морфема "*тиа*"), а вместо 5-го и 6-го атомов - атомы азота **N** (ему соответствует морфема "*аза*", а умножающая приставка "*ди*" показывает, что таких атомов два). Составной префикс "*3-метил-*" описывает боковую цепь, исходящую из 3-й вершины базисного углеродного каркаса.

Из приведённого примера явствует, что "Номенклатурный Анализатор" воспринимает вводимые названия химических веществ как сложные слова, вложенный смысл которых восстанавливается в процессе их дробления на элементарные фрагменты, в соответствии с грамматикой языка номенклатуры.

Описание работы "Номенклатурного Анализатора"

Программа "Номенклатурный Анализатор" представляет собой Windows-приложение, основной задачей которого является, как уже было сказано, построение структуры органического химического соединения по его названию. Название соединения вводится пользователем.

Представляемая версия Анализатора способна обрабатывать названия соединений следующих видов:

1) Ациклические (алифатические) углеводороды с нормальной или разветвлённой цепью:

а) предельные (насыщенные) углеводороды (т.н. *алканы*), т.е. соединения, в которых атомы углерода соединены только простыми (одинарными) связями C - C;

б) непредельные (ненасыщенные) углеводороды, т.е. соединения, в которых имеется одна пара углеродных атомов, соединённых кратными связями: двойными C = C (т.н. *алкены*) или тройными C ≡ C (т.н. *алкины*);

в) соединения, содержащие две, три и более двойные связи (т.н. *алкадиены*, *алкатриены* и т.д.), и, аналогично, соединения, содержащие две, три и более тройные связи (т.н. *алкадиины*, *алкатриины* и т.д.);

г) соединения, содержащие и двойные и тройные связи одновремен-но (т.н. *енины*);

2) Простейшие моноциклические соединения (как с боковыми цепями, так и без них). Сюда входят т.н. *циклоалканы*, *циклоалкены*, *циклоалкины*, *циклоалкаполиены*, *циклоалкаполиины*, *циклоенины*, а также *циклополиенполиины*;

3) Важнейшие классы органических соединений:

а) одно- и многоатомные спирты;

б) простые эфиры;

в) альдегиды;

г) кетоны;

д) карбоновые и поликарбоновые кислоты;

е) сложные эфиры;

ж) некоторые галогенопроизводные (**-Cl**, **-Br**, **-F**, **-I**);

з) соединения, включающие некоторые азотсодержащие группы (*амино*, *нитро*);

4) Ациклические и моноциклические углеводороды, отдельные атомы углеродной цепи в которых замещены гетероатомами. Сюда относятся соединения, названные по "а"-номенклатуре.

На данном этапе алгоритм применим к соединениям, длина наибольшей цепи которых насчитывает до 80 вершин.

Суть алгоритма состоит в том, чтобы адекватно разделить введённое пользователем название на составные части (морфемы), а затем, используя приписанную этим морфемам стандартную химическую информацию и опираясь на их взаимное расположение, скомпилировать единую структуру всего соединения, согласно заложенным в память алгоритма синтаксическим номенклатурным правилам сочетаемости морфем.

По завершении работы алгоритма на экран выводятся сведения о структуре обработанного соединения, представленные в следующем виде:

1) Количество вершин (т.е. атомов углерода, либо заменяющих их других элементов);

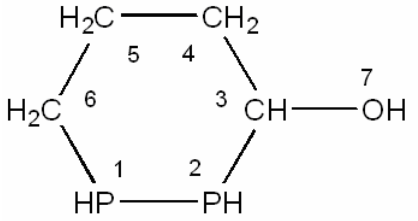
2) Перечень пронумерованных вершин (нумерация определяется алго-ритмом);

3) Общее число связей между вершинами данного соединения (причём двойные и тройные связи учитываются наравне с одинарными);

4) Перечень всех связей, для каждой из которых указываются номера соединяемых ею вершин и индекс, показывающий кратность связи между двумя этими вершинами. (Для одинарной связи индекс будет равен единице, для двойной - двум, для тройной - трём.)

Аналогичная информация выводится в специальный файл, в стандартном mol-формате, предназначенный для использования существующими на сегодняшний день отображающими программами, например - визуализатором HyperChem.

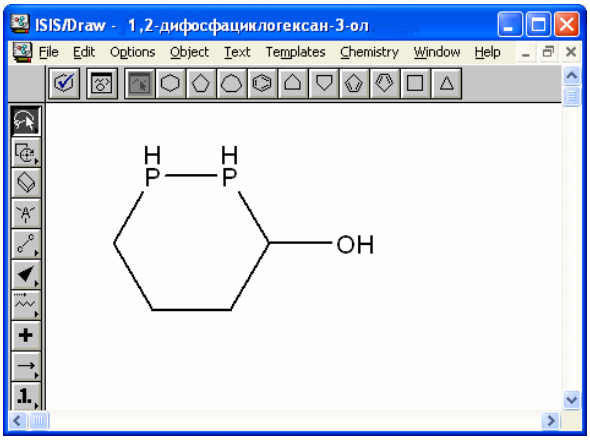
Так, при вводе названия "*1,2-дифосфациклогексан-3-ол*", которому соответствует структура



алгоритм выдаст следующий результат:

"1,2-дифосфациклогексан-3-ол
7 7
1 - PH 1-2 1
2 - PH 2-3 1
3 - CH 3-4 1
4 - CH2 4-5 1
5 - CH2 5-6 1
6 - CH2 1-6 1
7 - OH 3-7 1"

На основе этой информации алгоритм генерирует mol-файл. Результат отображения этого mol-файла графическим визуализатором ISIS/Draw можно видеть на рисунке:



Описание Анализатора будет неполным, если не упомянуть о встроенных в программу возможностях работы со словарём химических морфем.

Словарь является основной базой данных, на которой строится работа всей программы. Словарь содержит полный набор воспринимаемых алгоритмом морфем. Морфемы эти разбиты на несколько классов, что связано с различной их ролью при построении из них номенклатурного названия. После каждой морфемы следует соответствующая ей химическая информация, представленная в удобной для восприятия алгоритмом форме.

В диалоговом обеспечении Анализатора предусмотрены функции пополнения словаря, удаления из него элементов, сортировки его по классам морфем.

Перспективы задачи

Как следует из вышеизложенного, "Номенклатурный Анализатор" является программой, открытой для дополнений и доработок. Можно выделить основные направления дальнейшего развития Анализатора.

Это, прежде всего, расширение поля обрабатываемых названий с помощью пополнения словаря, введения новых классов морфем и дополнительных синтаксических правил сочетаемости. Особую актуальность имеет задача внесения в словарь программы тривиальных названий для ароматических и гетероциклических конденсированных соединений.

Кроме того, остаётся пока не решённая проблема с отображением в молекулярном графе стереохимических параметров соединений.

Необходимо также усовершенствовать графический аспект задачи, так как на настоящий момент программа не располагает встроенным визуализатором, вследствие чего приходится прибегать к "услугам" других программ.

Литература

1. *Уткина Е.А. Программа перевода названий химических соединений в систематической номенклатуре в молекулярные графы (для некоторых важных классов органических соединений) // НТИ. Серия 2. Информационные процессы и системы. - 2000. - № 3. - С. 24-36.*
2. *Номенклатура органических соединений. Справочник химика. Дополнительный том. // Изд-во "Химия", Ленинградское отделение, 1968.*
3. *Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F and H. Oxford, Pergamon Press, 1979.*
4. *Chemical Abstracts. Index Guide. Chemical Abstracts Service. The American Chemical Society, 1992.*
5. *Терентьев А.П. и др. Номенклатура органических соединений. - М.: Изд-во АН СССР, 1955.*
6. *Ланглебен М.М. О синтезе названий химических соединений // НТИ. - 1965. - № 10. - С. 18-24.*
7. *Ланглебен М.М. К лингвистическому описанию номенклатуры органической химии // НТИ. - 1967. - № 1. - С. 13-22.*
8. *Ланглебен М.М. Опыт приспособления лингвистических понятий и лингвистической терминологии к описанию искусственного языка // Информационные поисковые системы и автоматическая обработка научно-технической информации. - 1967. - С. 170-224.*
9. *Ланглебен М.М. Структура номинативных сочетаний в специальном фрагменте русского химического языка: Диссертация кандидата химических наук. // М.: ВИНИТИ, 1970. - 257 с.*
10. *Цукерман А.М. Номенклатура органических соединений и номенклатурный перевод. // М., 1966. - 253 с.*