УДК 003.6:544.131

Л. А. Григорян, В. В. Бондарь, И. Б. Немировская

Программа перевода систематических названий химических соединений в молекулярные графы (расширение на заменительную номенклатуру)

Разработана новая версия программы "Номенклатурный Анализатор", переводящей вводимые пользователем названия химических соединений, данные в систематической номенклатуре ИЮПАК, в молекулярные графы, представленные стандартным образом. Модернизированная и расширенная версия программы обеспечивает восприятие Анализатором так называемой заменительной, или "а"-номенклатуры в ациклических и моноциклических соединениях, что является первым шагом на пути к обработке названий гетероциклических соединений.

ПРЕДЫСТОРИЯ ВОПРОСА

Из многочисленных химических номенклатур [1] наиболее известны номенклатура Международного союза теоретической и прикладной химии (International Union of Pure and Applied Chemistry, ИЮПАК) [2] и номенклатура Американского химического общества (American Chemical Society) — номенклатура CAS [3]. Русскоязычным аналогом 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.

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 отражающее эту структуру. Следовательно, принятые в номенклатуре названия соединений должны делиться на отдельные компоненты, обладающие определённой информацией. Имея полный набор таких компонентов и руководствуясь номенклатурными правилами обращения с ними, можно выстроить из них, как из кубиков, любое верное химическое название.

Отдельную проблему составляет ситуация с так называемыми "тривиальными" названиями, т. е. с названиями, сложившимися исторически или привычно используемыми. 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.

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, так как 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 [4–7]. Ею же создана грамматика в виде порождающей модели синтаксиса номенклатуры органических соединений. Алгоритм номенклатурного перевода, использованный в "Номенклатурном Анализаторе" Е. А. Уткиной [8], разработан А. М. Цукерманом [9].

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 очень много времени, современный компьютер в состоянии осуществить за доли секунды (естественно, если задача алгоритмизирована). 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, структуру которых распознать 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 [2] 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.

ЗАМЕНИТЕЛЬНАЯ ("А"—) НОМЕНКЛАТУРА

Для составления названий определённого класса химических соединений применяются правила, как общие для всей номенклатуры, так и частные, пригодные только для данного класса соединений [3]. Рассмотрим особенности построения названий органических химических соединений, у которых в некоторых вершинах вместо атомов углерода находятся атомы других элементов, таких как кислород, азот, фосфор и т. п. Список номенклатурных правил, определяющих порядок присвоения названий соединениям такого типа, называется заменительной, или "а"-номенклатурой. Таким образом, заменительная номенклатура является как бы "подноменклатурой" общей номенклатуры.

Заменительная номенклатура базируется на принципе замены отдельных атомов углерода в химическом соединении другими атомами (их называют *гетероатомами)*; этому принципу она и обязана своим названием.

Эта номенклатура применяется прежде всего в:

1) Алифатических (т. е. ациклических) углеводородах, если в какой-либо из цепей имеется три или более гетероатомов. (Если же гетероатомов меньше, то соединение получает имя по другим принципам.)

2) Моноциклических углеводородах, если хотя бы одна из вершин заменена гетероатомом.

3) Радикалах всех тех соединений, для которых употребима эта номенклатура (радикалами называются фрагменты молекул, имеющие одну или более исходящих связей).

Кроме того, заменительная номенклатура используется и в некоторых других классах органических соединений, пока ещё не охваченных программой "Номенклатурный Анализатор".

Список гетероатомов, используемых в заменительной номенклатуре, правилами ИЮПАК задан однозначно (см. табл.):

Элемент	Символ	Валентность	Морфема
Кислород	O	II	окса
Сера	S	II	тиа
Селен	Se	II	селена
Теллур	Te	II	теллура
Азот	N	III	аза
Фосфор	P	III	фосфа
Мышьяк	As	III	арса
Сурьма	Sb	III	стиба
Висмут	Bi	III	бисма
Кремний	Si	IV	сила
Германий	Ge	IV	герма
Олово	Sn	IV	станна
Свинец	Pb	IV	плюмба
Бор	B	III	бора
Ртуть	Hg	II	меркура

Важен и порядок, в котором приводятся эти элементы.

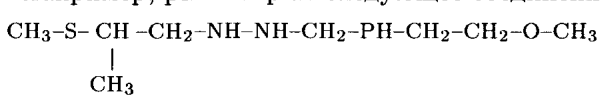
Как видно из табл. 1, каждому элементу данного списка соответствует определённая морфема, используемая заменительной номенклатурой для построения названий: для кислорода это "*окса*", для серы — "*тиа*", для азота — "*аза*", для фосфора — "*фосфа*" и т. п. Каждая из этих морфем имеет окончание "-*а*"; отсюда второе название заменительной номенклатуры — "*а*"-номенклатура.

Заменительная номенклатура устанавливает следующие правила построения названий ациклических органических соединений с гетероатомами. Самую длинную цепь, состоящую из атомов углерода и гетероатомов, называют как ациклический углеводород такой же длины без гетероатомов. Гетероатомы, входящие в эту цепь, обозначают указанными морфемами "*окса*", "*тиа*", "*аза*" и т. д. в порядке, определяемом вышеприведённым списком гетероатомов, с локантами, указывающими их положение в цепи. Применяются стандартные правила для использования умножающих приставок и выражения ненасыщенности. Цепь нумеруют от одного конца к другому, исключая концевые гетероатомы, входящие в функциональные группы. Порядок нумерации цепи определяется в первую очередь положением главной функциональной группы, которая должна получить по возможности наименьший номер. При выполнении этого условия и наличии альтернативы, нумерация осуществляется так,

чтобы любой ближайший к началу отсчёта гетероатом получил наименьший номер. При сохранении возможности выбора после выполнения предыдущих требований цепь нумеруют так, чтобы наименьший номер получил гетероатом, находящийся в начале указанного списка гетероатомов.

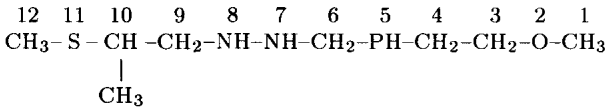
Таким образом, название цепи, содержащей гетероатомы, будет отличаться от названия такой же цепи, не содержащей гетероатомов, специальным *гетеропрефиксом* (в котором будет храниться информация о положении, количестве и роде гетероатомов) и, возможно, направлением нумерации. Если в цепи есть ответвления, то префиксы, описывающие их, будут поставлены в названии левее гетеропрефикса.

Например, рассмотрим следующее соединение:



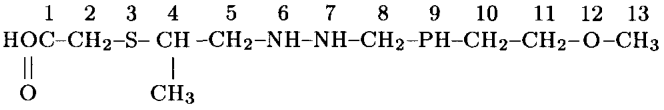
В нём сразу видна главная цепь, состоящая из 12 вершин, и один боковой метильный заместитель. Такое же соединение, в котором вместо гетероатомов стояли бы группы "**-CH_2-**", называлось бы "*3-метилдодекан*", где локант "*3*" указывал бы на положение боковой цепи относительно главной цепи, компоненты "*мет*" и "*ил*" описывали бы боковую цепь, состоящую из одной вершины, морфема "*додек*" называла бы число атомов в главной цепи (12), а суффикс "*ан*" говорил бы о насыщенности соединения. Нумерация главной цепи начиналась бы слева, чтобы метильный заместитель получил минимальный номер (3, а не 10). Но, поскольку в соединении присутствуют гетероатомы, следует по-новому решить вопрос о нумерации в цепи. Минимальный номер должен получить гетероатом, ближайший к концу цепи.

Однако на равном расстоянии от концов цепи находятся два гетероатома: слева — атом серы, справа — атом кислорода. Чтобы выбрать из них один, обратимся к приведённому выше гетеросписку и увидим, что кислород расположен в нём раньше серы. Следовательно, минимальный номер должен получить атом кислорода. Поэтому, нумерация в главной цепи пойдёт справа налево:



В результате получен гетеропрефикс: "*2-окса-11-тиа-7,8-диаза-5-фосфа*". Локант "*2*" указывает положение кислорода, "*11*" — серы, "*7*" и "*8*" — азота, а "*5*" — фосфора. Умножающая приставка "*да*" напоминает, что атомов азота в цепи имеется два. Морфемы "*окса*", "*тиа*", "*аза*" и "*фосфа*" определяют род гетероатомов; порядок следования этих морфем в названии определяется вышеприведённым гетеросписком. Кроме гетеропрефикса имеется также префикс, описывающий боковую ветвь, — при новой нумерации это будет "*10-метил*". В главной цепи 12 атомов, значит, основа названия соединения — "*додекан*". Собрав всё вместе, получим: "*10-метил-2-окса-11-тиа-7,8-диаза-5-фосфадодекан*".

Стоит заметить, что если к крайней левой вершине добавить карбоксильную группу "**-COOH**", то направление нумерации снова придётся менять, потому что в ациклических соединениях функциональным группам отдаётся приоритет перед всеми прочими заместителями и гетероатомами:



Кроме того, длина главной цепи увеличится на одну вершину, за счёт атома углерода, принадлежащего карбоксильной группе. Соответственно, такое соединение получит название "*4-метил-12-окса-3-тиа-6,7-диаза-9-фосфатридекановая кислота*".

Аналогично строится название и для моноциклических органических соединений с гетероатомами. Единственное отличие — гетероатомам в этом случае отдаётся приоритет над функциональной группой, а не наоборот, как было справедливо для ациклических соединений.

Рассмотрим пример циклического соединения с одной функциональной группой (рис. 1). Каркас (главная цепь) данного соединения — шестичленный моноцикл. При отсутствии в нём гетероатомов он назывался бы "*циклогексанол*". Морфема "*цикло*" здесь указывает на цикличность соединения, корень "*гекс*" означает, что цепь состоит из шести вершин, суффикс "*ан*" свидетельствует о насыщенности соединения, а "*ол*" соответствует имеющейся спиртовой функциональной группе "**-OH**".

Однако в данном случае две углеродные вершины замещены атомами фосфора. Поэтому в названии будет присутствовать гетеропрефикс. Чтобы правильно назвать это соединение, надо прежде всего определить, от какой вершины и в какую сторону пойдёт нумерация. Наименьшие локанты (в данном случае — "*1*" и "*2*") должны быть у вершин, замещённых гетероатомами. Далее нумерация пойдёт против часовой стрелки — тогда функциональная группа будет присоединена к 3-й вершине (а не к 6-й, как было бы в противном случае).

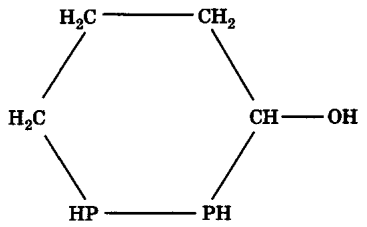


Рис. 1

Таким образом, положение локантов определено, и можно составить гетеропрефикс: "*1,2-дифосфа*". К суффиксу в названии моноцикла тоже придётся добавить локант — получится "*циклогексан-3-ол*". Соединив оба этих фрагмента названия, получаем систематическое наименование соединения: "*1,2-дифосфациклогексан-3-ол*" (рис. 2).

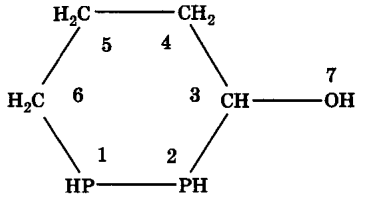


Рис. 2

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) одноатомные 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 50 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.

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 алгоритмом);

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 рассмотренная выше структура (рис. 3)

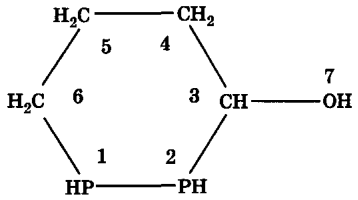


Рис. 3

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 рис. 4:

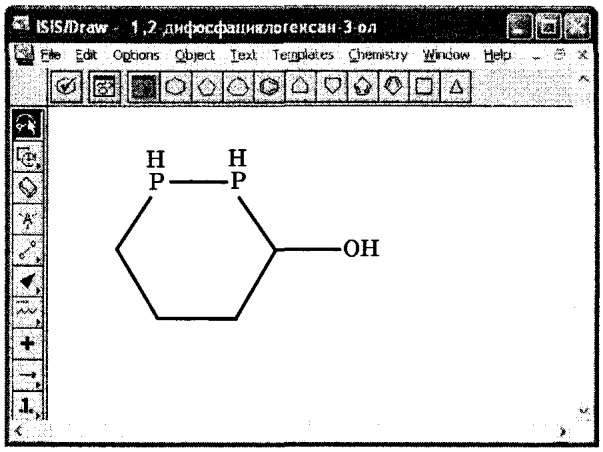


Рис. 4

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.

КОРРЕКЦИЯ АЛГОРИТМА "НОМЕНКЛАТУРНОГО АНАЛИЗАТОРА"

Модернизация исходной версии "Номенклатурного Анализатора" в целях увеличения числа классов обрабатываемых им соединений включала следующие этапы:

- Модернизация набора типов морфем, используемых алгоритмом, введение особых Hetero-морфем;
- Пополнение словаря программы рядом новых морфем, принадлежащих ко вновь введённому типу;
- Аналитическая обработка вводимого пользователем названия химического соединения на предмет поиска и вычленения т. н. гетеропрефиксов, состоящих из Hetero-морфем, соответствующих им локантов, умножающих приставок и служебных знаков;
- Предварительное соотнесение найденных гетеропрефиксов с соответствующими им углеродными цепями создаваемого программой единого графа химического соединения;
- Структурирование содержащейся в найденных гетеропрефиксах химической информации;
- Интеграция этой информации в порождаемый алгоритмом единый граф химического соединения путём замены соответствующих углеродных вершин гетероатомами;
- Автоматическая проверка верной расстановки валентностей всех вершин по всем цепям единого графа химического соединения;
- Выявление некоторых возможных ошибок пользователя, допущенных им при введении номенклатурного названия химического соединения.

Для внесения этих изменений программа была существенно переработана, снабжена новыми функциями-обработчиками, в алгоритм внесены значительные дополнения и уточнения программистского характера. Кроме того, словарь Анализатора был пополнен морфемами, описывающими углеродные каркасы длиной от 31 до 50 вершин (прежняя версия программы ограничивала максимальную длину цепи тридцатью вершинами).

Проведено масштабное тестирование работы программы на экспериментальном массиве, включающем названия химических соединений, принадлежащих ко всем описанным классам.

ПЕРСПЕКТИВЫ ЗАДАЧИ

Как следует из изложенного, "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 или введения новых классов морфем. Особую актуальность имеет задача внесения в словарь программы тривиальных названий для ароматических и гетероциклических конденсированных соединений.

Кроме того, пока остаётся не решённой 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 этого приходится прибегать к "услугам" других программ.

СПИСОК ЛИТЕРАТУРЫ

1. Номенклатура органических соединений. Справочник химика. Дополнительный том. Изд-во "Химия", Ленингр. отд., 1968.
2. Nomenclature of Organic Chemistry. Sections A, B, C, D, E, F and H. Oxford, Pergamon Press, 1979.
3. Chemical Abstracts. Index Guide. Chemical Abstracts Service. The Amer. Chem. Society, 1992.
4. Ланглебен М. М. О синтезе названий химических соединений // НТИ.— 1965.— № 10.— С. 18–24.
5. Ланглебен М. М. К лингвистическому описанию номенклатуры органической химии // НТИ. Сер. 2.— 1967. — № 1.— С. 13–22.
6. 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.
7. Ланглебен М. М. Структура номинативных сочетаний в специальном фрагменте русского химического языка: Диссертация кандидата химических наук.— М.: ВИНИТИ, 1970.— 257 с.
8. Уткина Е. А. Программа перевода названий химических соединений в систематической номенклатуре в молекулярные графы (для некоторых важных классов органических соединений) // НТИ. Сер. 2.— 2000.— № 3.— С. 24–36.
9. Цукерман А. М. Номенклатура органических соединений и номенклатурный перевод.— М., 1966.— 253 с.

Материал поступил в редакцию 14.12.05