ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [003.6:544.131]:004.9

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

Программа перевода систематических названий химических соединений в молекулярные графы (расширение на класс ароматических соединений)

Разработана новая версия программы "Номенклатурный Анализатор", переводящей систематические названия химических соединений в молекулярные графы. Программа является развитием проекта проф. В. К. Финна, основу которого заложила в 1999 г. программист Е. А. Уткина [1]. Созданная ею исходная версия "Номенклатурного Анализатора" была способна обрабатывать названия, относящиеся лишь к некоторым важным классам органических соединений. Программа, представляемая в данной статье, обрабатывает новый класс — соединения, традиционно именуемые ароматическими, что является очередным шагом вперёд на пути к охвату всей русскоязычной химической номенклатуры.

ХИМИЧЕСКАЯ НОМЕНКЛАТУРА

Номенклатура названий химических соединений — главная и незаменимая составляющая языка химической науки. Трудно переоценить значимость химической номенклатуры, именно номенклатура отвечает за присвоение названий десяткам миллионов химических соединений. Без точных, чётких, прозрачных и продуманных номенклатурных правил в химии воцарился бы хаос, подобный тому, что имел место в биологии до появления классификации Карла Линнея.

В настоящее время в мире бытует несколько "соперничающих" химических номенклатур международного класса [2]. Среди них 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.

Основной принцип химической номенклатуры, установившийся ещё в XIX веке, заключается в том, что название химического соединения должно максимально отражать его структуру. Те компоненты, из которых строится наименование химического вещества, должны определённым образом соответствовать компонентам структурного графа этого вещества — цепочкам атомов, связям различной кратности, стыковым позициям подграфов, нестандартным химическим элементам, присутствующим в соединении и т. п.

Кроме того, поскольку многообразие химических соединений чрезвычайно велико, оказалось удобно сгруппировать эти соединения по различным классам, в зависимости от их свойств, строения и сложности. Подобный подход потребовал известной гибкости от химической номенклатуры, в которой появились соответствующие разделы и подразделы (как бы встроенные подноменклатуры), описывающие особенности построения наименований для отдельных классов химических соединений.

В работах [5, 6] кратко рассмотрены основные положения соответственно заменительной номенклатуры и номенклатуры Ганча-Видмана, представляющие собой немаловажные разделы общей химической номенклатуры ИЮПАК [3]. В настоящей работе рассматривается ещё один раздел номенклатуры ИЮПАК, известный под названием номенклатуры ароматических соединений.

ПРЕДПОСЫЛКИ К РАЗРАБОТКЕ ПРОГРАММЫ "НОМЕНКЛАТУРНЫЙ АНАЛИЗАТОР"

Как следует из вышеизложенного, между систематическим названием химического соединения и его структурой существует определённое соответствие. На практике этот принцип означает возможность по названию соединения восстановить его

структуру, равно как и по структуре получить систематическое название. Обе эти взаимнообратные процедуры осуществимы только при строгом применении установленных номенклатурой правил.

Естественно, если дело касается достаточно сложных структур или многоэлементных названий, осуществление подобного перевода вручную чревато разнообразными ошибками и погрешностями, искажающими смысл обрабатываемой информации. Кроме того, работы такого рода являются чрезвычайно времяёмкими и, помимо всего прочего, для них требуются специально обученные специалисты, детально разбирающиеся в вопросах номенклатуры (важно отметить, что полный свод номенклатурных правил с примерами их применения насчитывает несколько томов и постоянно обновляется).

Тем не менее, такого рода задачи, возникающие, например, при подготовке к печати реферативных журналов или патентов на изобретения, до последнего времени решались практически вручную, без использования современных компьютерных систем, способных автоматизировать рутинный процесс обработки информации, минимизировать вероятность возникновения ошибок и визуализировать полученные результаты.

Программа "Номенклатурный Анализатор" позволяет устранить это отставание и значительно облегчить труд научных работников, которые занимаются построением структур химических соединений по названию.

Обратная сторона задачи (т. е. восстановление названия по имеющейся структуре), алгоритмически менее сложная, была решена сравнительно недавно сразу несколькими различными организациями-разработчиками программного обеспечения для нужд химической науки. Анализатор же, ответственный за перевод "структура → название", является беспрецедентной разработкой на пространстве бывшего СССР. За рубежом уже имеются программы, выполняющие сходную функцию (например, немецкий пакет AutoNom, канадский номенклатор ACD/Labs, или кембриджский пакет ChemOffice), но ни одна из них не поддерживает русскоязычную номенклатуру.

Теоретическую основу программы "Номенклатурный Анализатор" заложили работы М. М. Лангглебен [7–10] и А. М. Цукермана [11], впервые рассмотревших химическую номенклатуру как искусственный язык, допускающий чёткую формализацию и обладающий развитой порождающей грамматикой. А. М. Цукерман [11] также известен как разработчик первых алгоритмов, положенных в основу Анализатора. Практическую реализацию этим начинаниям дала программист Е. А. Уткина, создавшая в 1999 г. (под руководством проф. В. К. Финна) исходную пробную версию Анализатора. Однако та версия программы была способна обрабатывать лишь самые простые классы химических соединений, а потому нуждалась в существенной доработке и качественном развитии.

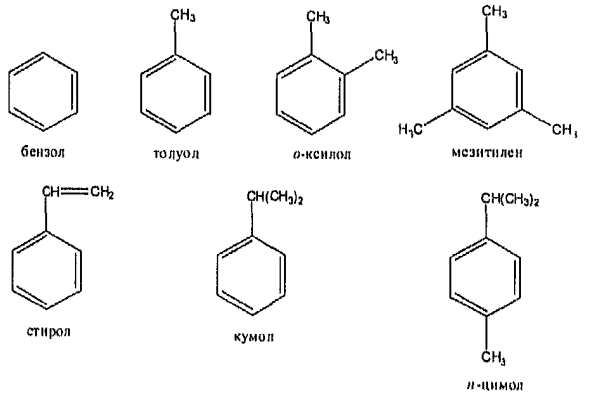
Версия программы, представляемая в данной статье, существенно отличается от предыдущих [1, 5, 6] и позволяет обрабатывать новые классы химических соединений, в том числе важнейшую группу ароматических соединений.

НОМЕНКЛАТУРА АРОМАТИЧЕСКИХ СОЕДИНЕНИЙ

Согласно принятому определению из области органической химии, к ароматическим соединениям относятся плоские циклические соединения, атомы которых вносят в π-электронную систему молекулы $4n+2$ p-электронов (n может принимать целочисленные значения 1, 2, 3, 4 и далее). Молекулы таких соединений стабилизированы в силу значительной величины энергии делокализации π-электронов. Для соединений этого ряда характерны реакции замещения, а не присоединения, в отличие от обычных сопряжённых полиенов. Первый представитель ароматических соединений, молекула которого имеет плоский циклический каркас из шести атомов углерода с тремя чередующимися двойными связями, называется не "*циклогекса-1,3,5-триен*", а "бензол".

Все ароматические соединения имеют тривиальные (т. е. не систематические, а сложившиеся исторически) или полутривиальные названия, такие как, например, "*стирол*", "*толуол*", "*нафталин*" и др.

К моноциклическим ароматическим соединениям относятся, прежде всего, *бензол* и шесть его гомологов. Для них номенклатура ИЮПАК устанавливает [12] следующие тривиальные названия:

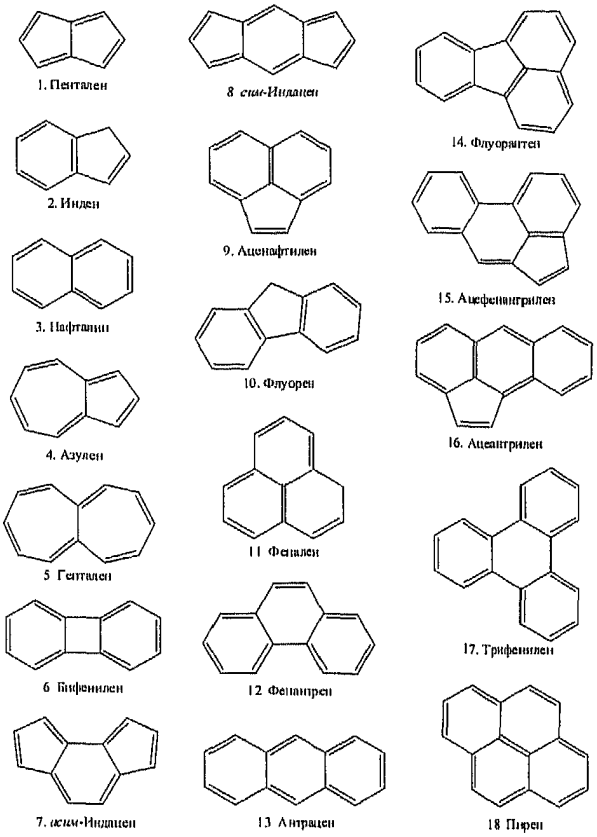


бензол, толуол, *o*-ксилол, мезитилен, стирол, кумол, *n*-цимол

Простейшие гомологи бензола

Ароматика как класс химических соединений занимает непревзойдённое по важности место среди всех видов органических химических веществ. В подавляющем большинстве химических соединений присутствуют фрагменты, относящиеся к классу ароматических соединений.

Номенклатура названий ароматических соединений, строго говоря, систематической не является, поскольку опирается не на классические морфемы алифатики (хотя и они тоже часто присутствуют в названиях), а на простые и составные тривиальные компоненты, которые лишь с некоторой натяжкой можно считать морфемами. Поэтому, в отсутствие чёткой порождающей системы названий, основные ароматические углеводороды заданы списком:



Начальные 18 полициклических ароматических соединений, для которых правилами ИЮПАК установлены тривиальные названия

С лингвистической точки зрения в этих названиях можно выделить суффикс "*ен*", свидетельствующий (для данного раздела номенклатуры) о наличии максимального числа чередующихся двойных связей (исключение составляет наименование "*нафталин*", сохраняемое в силу сложившихся традиций, -- правильнее было бы использовать термин "*нафтален*"). Однако для программистского представления задачи удобнее считать приведённые выше названия неделимыми.

Легко видеть, что базовым элементом ароматических соединений является бензольное кольцо. Различные комбинации этих колец (конденсирование), собственно и составляют основную массу ароматических соединений.

Нормальным состоянием для ароматического соединения является ненасыщенность, что выражается наличием максимально возможного количества некумулированных (т. е. не примыкающих друг к другу) двойных связей. При этом в углеродных вершинах соединения содержится, как правило, по одному атому *водорода* (в местах конденсации циклов атомы *водорода* отсутствуют). Однако нередки и случаи частичного или полного гидрирования вещества. Тогда в соответствующие вершины соединения добавляется ещё один атом *водорода*, а двойные связи заменяются одинарными. Таково, например, вещество *1,2,3,4-Тетрагидронафталин* (для удобства называемое так же "*тетралин*") (см. рис. 1):

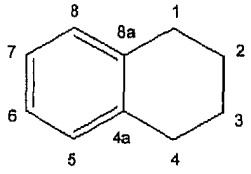


Рис. 1. 1,2,3,4-Тетрагидронафталин

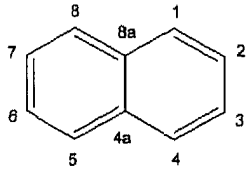


Рис. 2. Нафталин

Как видно из рис. 1, характерные для обычного *нафталина* (рис. 2) двойные связи между 1-й и 2-й, а также между 3-й и 4-й вершинами — отсутствуют, а в указанных вершинах находятся по два атома *водорода*.

Номенклатурно гидрирование описывается точно так же, как это было в случае соединений Ганча-Видмана [6].

Ещё одна номенклатурная особенность, справедливая для конденсированных ароматических соединений, связана с нумерацией вершин. Если в рассмотренных прежде [1, 5, 6] классах химических соединений допускались только обычные числовые локанты *1, 2, 3* и т. д., то здесь вершины, расположенные на стыках колец, обозначаются число-буквенными локантами, такими как, например, *3a* или *5b*. Нумерация ведётся последовательно по периметру графического отображения структуры соединения, по часовой стрелке, начиная с первого несмежного атома в правом верхнем цикле. В качестве показательного примера можно привести *хризен* (см. рис. 3) и *бифенилен* (см. рис. 4).

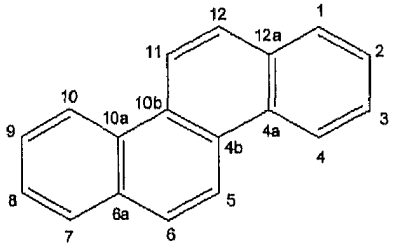


Рис. 3. Хризен

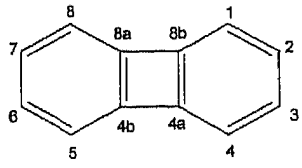


Рис. 4. Бифенилен

Исключение в нумерации представляет *антрацен* (см. рис. 5):

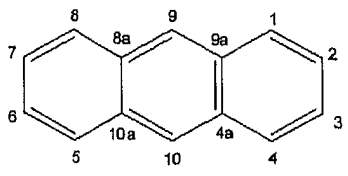


Рис. 5. Антрацен

Кроме того, непосредственно физико-химическими свойствами ароматических соединений обусловлено немаловажное их свойство. Дело в том, что двойные связи в бензольном кольце распространены равномерно, то есть как бы "размазаны" в равной мере между всеми вершинами кольца. Поэтому, несмотря на графическое различие между рисунками 6а и 6б, оба они соответствуют одному и тому же соединению — *бензолу*.

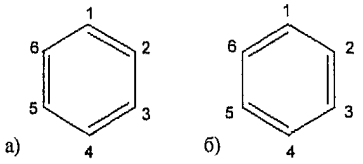


Рис. 6. Бензол

Выбор одного из двух представленных вариантов расстановки двойных связей — не принципиален. И эта особенность справедлива для любого, сколь угодно сложного, сочетания бензольных колец — т. е., фактически, для всех ароматических соединений.

Так, например, *трифенилену* в равной степени соответствуют как рисунок 7а, так и рисунок 7б.

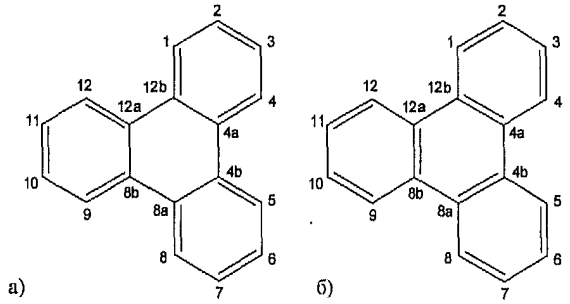


Рис. 7. Трифенилен

Часто для удобства обозначения двойные связи на графах, относящихся к бензольным кольцам, вообще не прорисовывают, а заменяют одной лишь вписанной окружностью (см. рис. 8):

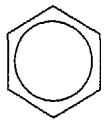


Рис. 8. Бензол

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 одновременно (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. Ациклические и моноциклические углеводороды, отдельные атомы углеродной цепи в которых замещены гетероатомами. Сюда относятся соединения, названные по "а"-номенклатуре;
4. Ароматические углеводороды:
   а) предельные;
   б) непредельные;
   в) с боковыми заместителями;
   г) с основными функциональными группами;
5. Моноциклические углеводороды, содержащие гетероатомы, названные по номенклатуре Ганча-Видмана;
6. Важнейшие классы органических соединений:
   а) одно- и многоатомные спирты;
   б) простые эфиры;
   в) сложные эфиры;
   г) кетоны;
   д) альдегиды;
   е) карбоновые и поликарбоновые кислоты;
   ж) некоторые галогенопроизводные (в том числе —Cl, —Br, —F, — I);
   з) соединения, включающие некоторые азотсодержащие группы (*амино, нитро*).

Подключение к множеству обрабатываемых Анализатором названий ароматических соединений с нестандартной гидрированностью отдельных вершин находится в стадии тестирования.

На данном этапе алгоритм применим к соединениям, длина наибольшей цепи которых насчитывает до 100 вершин.

Суть алгоритма состоит в том, чтобы правильно разделить введённое пользователем название химического соединения на простейшие 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 и может не совпадать с номенклатурной);

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 "*4-этил-8,9-дигидроперилен*", которому соответствует структура (рис. 9)

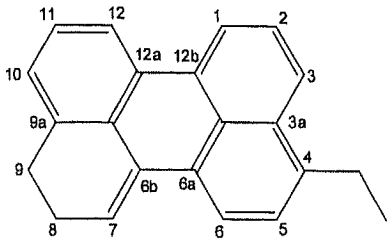


Рис. 9

алгоритм выдаст следующий результат:

"4-этил-8,9-дигидроперилен

22	26
1 —CH	1–2 1
2—CH	2–3 2
3 -CH	3–4 1
4- - C	4–5 1
5 - -C	5–6 2
6- CH	6–7 1
7 -CH	7–9 2
8 -C	9–8 1
9- -C	8–10 2
10- -CH	10–11 1
11 - CH	11–12 1
12 -CH	12–13 1
13 C	13–14 1
14- CH	14–15 2
15 -CH	15–16 1
16 CH	16–19 2
17 C	19–18 1
18 C	18–1 2
19 -C	18 17 1
20 C	4 17 2
21 CH2	9–17 1
22 CH3	19–20 1
	13 20 2
	8–20 1
	5–21 1
	21–22 1"

На основе этой информации алгоритм генерирует mol-файл. Результат отображения этого mol-файла графическим визуализатором ISIS/Draw можно видеть на рис. 10.

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.

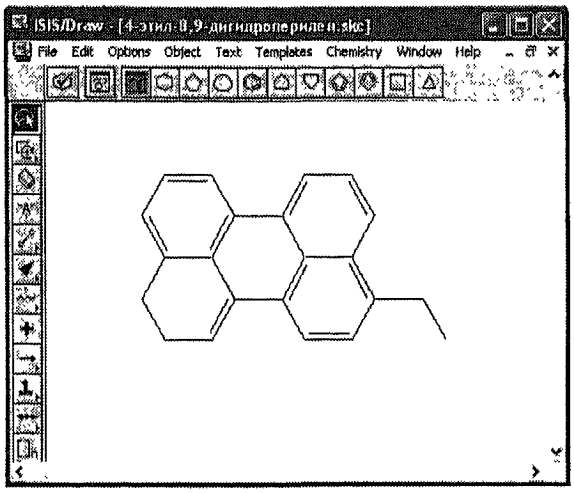


Рис. 10

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

КОРРЕКЦИЯ АЛГОРИТМА "НОМЕНКЛАТУРНОГО АНАЛИЗАТОРА"

Доработка предыдущей версии "Номенклатурного Анализатора" с целью подключения класса ароматических соединений к полю обрабатываемых им названий и структур состояла из следующих этапов.

- Пополнение опорного словаря программы морфемами, отвечающими за обозначение углеродного каркаса основных ароматических соединений.
- Пополнение набора контекстных правил алгоритма правилами обработки названий, включающих в себя морфемы, относящиеся к ароматическому классу соединений;
- Снабжение алгоритма программы функцией аналитической обработки вводимого пользователем названия химического соединения на предмет поиска и вычленения число-буквенных локантов;
- Снабжение алгоритма функцией пересчёта нумерации химического соединения с число-буквенной на сугубо числовую;
- Автоматическая проверка правильности переноса информации о гидрированности отдельных вершин химического соединения из соответствующих морфем в специальные массивы данных;
- Интеграция информации о повышенной гидрированности отдельных вершин химического соединения в порождаемый алгоритмом единый граф путём автоматического снятия двойных связей в соответствующих углеродных вершинах при одновременном увеличении свободных валентностей этих вершин;
- Координация между графом, относящимся к основному каркасу ароматического соединения, и графами боковых цепей соединения на предмет возможности их адекватной стыковки при создании единого графа химического соединения;

- 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.

При внесения этих изменений алгоритм программы был существенно скорректирован и дополнен. Потребовалось написание новых функций-обработчиков и процедур автоматической верификации. Кроме того, словарь Анализатора был пополнен некоторыми служебными морфемами, необходимыми для обработки отдельных неохваченных ранее соединений из числа уже подключённых к алгоритму классов химических веществ.

Программа прошла тестирование на специально подобранном экспериментальном массиве, включающем названия ароматических соединений, как с нормальной, так и с повышенной гидрированностью, как с боковыми цепями, так и без них.

ПЕРСПЕКТИВЫ И ЗАДАЧИ

"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 названий для гетероциклических конденсированных соединений, мостиковых соединений и ансамблей циклов.

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

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

СПИСОК ЛИТЕРАТУРЫ

1. 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.

2. Номенклатура органических соединений: Справочник химика Дополнительный том. Л . 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 Григорян Л. А., Бондарь В. В., Немировская И. Б. Программа перевода систематических названий химических соединений в молекулярные графы (расширение на заместительную номенклатуру) // НТИ. Сер. 2 2006. № 3. - С 21 25

6. Григорян Л. А. Программа перевода систематических названий химических соединений в молекулярные графы (расширение на номенклатуру Ганча-Видмана). В печати.

7. Ланглебен М. М. О синтезе названий химических соединений // НТИ.-- 1965.— № 10.— С. 18 24.

8. Ланглебен М М. К лингвистическому описанию номенклатуры органической химии // НТИ. 1967. — № 1. — С. 13–22.

9. 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.

10. Ланглебен М. М. Структура номинативных сочетаний в специальном фрагменте русского химического языка: Дисс. канд. хим. наук. М.: ВИНИТИ, 1970.— 257 с.

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

12. Кан Р., Дермер О. Введение в химическую номенклатуру — М.: Химия, 1983. 224 с.

Материал поступил в редакцию 01.06.06