Программный комплекс ГРИН для трехмерного представления органических молекул в естественных переменных

Атавин Е.Г., Фролова М.И.
Омский государственный университет
Омск, Россия
E-mail:Atavin@univer.omsk.su

Аннотация

Разрабатывается программный комплекс, позволяющий по систематическому (IUPAC) названию генерировать трехмерное компьютерное представление органических молекул в естественных переменных. При невозможности интерпретировать название или его фрагмент по правилам IUPAC делаются вывод об использовании несистематического термина и попытка найти его систематический аналог в базе данных несистематических названий молекул. При успешном анализе названия молекула представляется совокупностью атомных цепочек, для построения которых используются соответствующие математические алгоритмы [1] и база данных по эталонным значениям структурных параметров молекул.

1. Введение

Практически неограниченное количество органических молекул и широкое разнообразие характеризующих их свойств - химических, физических, биохимических, структурных, спектральных и т.д. предопределяют необходимость создания соответствующих баз данных. Данная работа посвящена моделированию трехмерного пространственного строения молекул, представляющего интерес не только с точки зрения их визуализации, но и как необходимый этап квантовохимических и молекулярно-механических расчетов, а также экспериментального определения строения молекул.

2. Описание молекулярной геометрии в декартовой и внутренней системах координат

Пространственное строение молекул принято описывать либо в декартовой системе координат, либо заданием значений структурных параметров — межъядерных расстояний, валентных углов и углов внутреннего вращения (естественных переменных).

Первый способ предполагает знание $3N$ декартовых координат N атомов, позволяет легко строить графическое изображение молекулы, вычислять значения всех структурных параметров и используется в большинстве современных программ квантовой механики, молекулярной механики и колебательной спектроскопии. Однако, произвол в выборе положения начала координат и ориентации координатных осей затрудняет сравнение результатов, полученных разными авторами. Кроме того, наличие у молекулы трех поступательных и трех вращательных степеней свободы приводит к появлению шести нулевых собственных значений у матрицы вторых производных энергии по координатам и к дополнительным осложнениям вычислительного характера. Наконец, само задание декартовых координат атомов — нетривиальная задача, поскольку они не являются справочными данными.

Альтернативное описание менее зависит от произвола исследователя, благодаря имеющимся эмпирическим закономерностям [2] в значениях структурных параметров. При оптимизации геометрии молекулы в естественных переменных можно упрощать задачу, фиксируя значения хорошо известных параметров или объединяя их в группы. Легко организовать поиск глобального минимума энергии, перебирая допустимые значения всех или некоторых параметров. При работе же с декартовыми координатами реализация этих возможностей сопряжена со значительными трудностями.

Однако, непосредственно по значениям структурных параметров невозможно в общем случае построить графическое изображение молекулы. Затруднительно также выполнять любые вычислительные операции с моделью молекулы, например, определять расстояния между удаленными атомами.

Таким образом, оба способа описания молекулярной структуры обладают рядом практически важных достоинств и весьма существенных недостатков. Совмещение достоинств достигается вычислением декартовых координат атомов по заданным структурным параметрам , что представляет собой в общем случае весьма громоздкую стереометрическую задачу. Алгоритмы ее решения значительно отличаются для топологически различных классов молекул и опубликованы [1].

3. Области использования структурной информации

Для создания эффективного программного комплекса необходимо конкретизировать цель моделирования пространственного строения молекул.

Использование структурной информации в приложениях можно свести к следущим вариантам.

1) Составление традиционного плоского изображения

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

молекул, необходимого, например, в полиграфии.

2) Создание пространственной модели молекулы для выбора ее проекции, наиболее удобной для обсуждения конкретной задачи, либо рассмотрения проблем стерического характера, например, биохимической комплементарности.

3) Расчет структурных, энергетических, спектроскопических характеристик, индексов реакционной способности и т.д., теоретическими методами. (Необходим ввод пространственного строения.)

4) Определение геометрического строения молекул экспериментальными методами. (Необходим ввод стартовой структуры.)

4. Способы ввода структурной информации

Во всех перечисленных вариантах требуется ввод информации о строении молекул, который можно осуществить следующими способами:

1) Ввод непосредственно декартовых координат атомов.

2) Графический ввод на экран плоского изображения молекулы с последующим переходом к трехмерному представлению.

3) Поиск в библиотеке заранее заготовленной структурной информации по ее названию, либо составление молекулы из заранее заготовленных фрагментов.

4) Написание подпрограммы, вычисляющей декартовы координаты атомов по значениям структурных параметров - межъядерных расстояний, валентных углов и углов внутреннего вращения (частный случай, часто используемый в спектроскопических и квантовохимических программах - аппарат Z-матриц).

5) Создание трехмерного изображения молекулы по ее названию.

Каждый из перечисленных способов задания информации о геометрическом строении молекулы имеет свои достоинства и недостатки.

1) Набор декартовых координат атомов, дополненный информацией о типах атомов и связей молекулы, является достаточным для построения трехмерного изображения молекулы. Он удобен для работы с квантовохимическими, спектроскопическими и пр. программами, а также как для использования в качестве стартового приближения в молекулярной механике и рентгеноструктурном анализе. Однако поскольку декартовы координаты атомов не обладают свойством транзитивности (например, атомы, занимающие симметрически эквивалентные позиции, имеют различные координаты), на их основе нельзя создать библиотеку структурных данных, и само их задание, вообще говоря, не является тривиальной задачей.

2) Рисование на экране плоского изображения с попыткой последующего перехода к трехмерному представлению очень удобно для малых и средних молекул и по существу является вариантом реализации предыдущего способа. Однако при вводе изображения на плоский экран происходит потеря части структурной информации, затрудняющая адекватное машинное представление конформеров сложных молекул. Причем по мере усложнения структуры вероятность случайного попадания в нужный конформер убывает. Легко убедиться в этом, попытавшись ввести таким способом (например, в методе молекулярной мехоники) конкретный конформер - твист-форму циклогексана или [3333]-циклододекан. Организовать же управляемый переход от плоского изображения к требуемому конформеру можно, по-видимому, лишь методом проб и ошибок.

Кроме того, наложение условий симметрии и других связей на структурные параметры, а также фиксация части из них для упрощения и ускорения расчетов труднореализуемы.

3) Поиск в библиотеке заранее заготовленной структурной информации по ее названию и составление молекулы из заранее заготовленных фрагментов активно используются для работы с накопленными структурными данными и в редакторах химических формул. Однако при отсутствии в базе данных нужной информации такой подход неприменим.

4) Написание подпрограммы, вычисляющей декартовы координаты атомов по значениям структурных параметров, является максимально исчерпывающим способом задания строения молекул, поскольку легко позволяет наложить на структуру условия симметрии и другие связи между значениями структурных параметров (например, ввести в расчет фиксированные значения малых разностей близких структурных параметров, трудно оцениваемых экспериментально, но сравнительно надежно рассчитываемых квантовохимически). При этом не вызывает трудностей при экспериментальном или теоретическом определении геометрии молекулы фиксация части хорошо известных структурных параметров (например, сравнительно мало меняющихся межьядерных расстояний, валентных углов, жестких фрагментов - фенильных радикалов и т.д.) для более быстрой и надежной оценки значений оставшихся параметров (обычно труднооцениваемых углов внутреннего вращения). Кроме того, в этом случае появляется возможность при нахождении пространственного строения молекул обращаться к базе данных по эталонным значениям межъядерных расстояний, валентных и торсионных углов, что позволяет прогнозировать пространственное строение молекул с экспериментально неисследованной структурой. Однако написание подобных геометрических подпрограмм не только весьма трудоемко и требует определенной квалификации, но и разрывает автоматизированный процесс обработки структурной информации.

5. Постановка задачи

Анализ достоинств и недостатков вышеприведенных способов ввода структурной информации приводит к выводу, что наиболее гибким и удобным может оказаться прямое "общение"с компьютером на естественном химическом языке.

Цель данной работы и состоит в создании программного комплекса, способного из названия молекулы по правилам номенклатуры IUPAC генерировать трехмерную модель молекулы во внутренних переменных (то есть с возможностью использования эталонных значений межъядерных расстояний, валентных и торсионных углов вместо нетабулированных декартовых координат атомов).

Для работы этого комплекса необходимо организовать две библиотеки: библиотеку значений эталонных межъядерных расстояний, валентных углов и углов внутреннего вращения для наиболее часто встречающихся типов атомов (близкую к библиотеке, имеющейся, например, в программах Молекулярной механики) и библиотеку, связывающую рациональные и тривиальные названия органических веществ с их систематическими анало-

гами.

6. Семантический анализ названия молекулы

Не касаясь основ систематической номенклатуры, напомним, что правила построения систематических названий молекул разработаны так, чтобы названия молекул были однозначно связаны с их химическим (топологическим) строением. Некоторое расширение этих правил для характеристики конформационных состояний линейных (задание конформации гош+, гош- или анти для каждой одинарной связи) и циклических (например, использование номенклатуры Гото-Даля [3]) молекул позволяет добиться взаимно однозначного соответствия названия и пространственного строения молекулы. Следовательно, семантический анализ систематического названия молекулы способен дать всю информацию, необходимую для компьютерного моделирования ее пространственного строения.

Для извлечения структурной информации полезно в IUPAC-названии молекулы можно выделить следующие морфемы:

КОРЕНЬ - часть названия, определяющая длину главной цепи (МЕТан, ЭТен, ПРОПин). Корень находится слева от ОКОНЧАНИЯ (см.ниже), КОЭФФИЦИЕНТА ПОВТОРЕНИЯ или СПИСКА, относящихся к ОКОНЧАНИЮ.

ПРОПан, БУТА-2,4-диен, БУТАдиен-2,4

Одно или несколько ОКОНЧАНИЙ - АН, ЕН, ИН, ОЛ и т.д., связанных с наличием функциональных групп или кратных связей и определяющих принадлежность химического соединения к определенному классу веществ.

пропАН, бутЕН-1, бутанОЛ-2, бут-1-ЕН-3-ОЛ

ПРЕФФИКСЫ - широкая группа морфем, определяющих соответствующие структурные особенности как главной цепи (ЦИКЛО, БИЦИКЛО, ТРИЦИКЛО,...), так и боковых цепей (ИЛ, ИЗО, ВТОР, ОКСИ, ОКСО, ИЛИДЕН, R, S,...). Во втором случае преффиксу предшествует новый корень, определяющий длину боковой цепи, возможно, со своими преффиксами и т.д.. Для удобства все корни, кроме первого, при наличии у них собственных преффиксов заключаются в скобки (эта дополнительная символика является избыточной, дублируя пару смысловых элементов имени "СПИСОК— "ИЛ").

5-(1-метилпропил)нонан

Окончания и преффиксы могут иметь КОЭФФИЦИЕНТЫ ПОВТОРЕНИЯ - ДИ, ТРИ, ТЕТРА и т.д..

5,5-ДИ(1,1,2,2-ТЕТРАметил)нонан

Перед преффиксами и окончаниями, а иногда и после окончений могут располагаться СПИСКИ - последовательность цифр, разделенных запятыми, количество которых определяется соответствующим коэффициентом повторения. Цифры списка определяют места присоединения боковых цепей.

В химической практике распространено использование также несистематических названий молекул, например, бензол, индол, нафталин и т.д. или радикалов - фенил, ацетил, бензоил и т.п.. Очевидно, что попытка интерпретации таких названий как слов IUPAC либо невозможна, либо приведет к неправильному результату. Последнее связано с возможной неоднозначноостью несистематического языка и должно контролироваться пользователем (например, техническое название 2,2,4-триметилпентана - "изооктан"по правилам IUPAC интерпретируется как 2-метилгептан).

Уточним, что буквосочетания "ол", "ил"и др. в подобных названиях не будут интерпретированы как соответствующие преффиксы, так как рядом с ними не будет найден соответствующий СПИСОК, либо КОРЕНЬ. Следует отметить, что некоторые названия, которые традиционно принято относить к систематическим, например, метанол, этанол, этен, этин, формально таковыми не являются по причине отсутствия СПИСКА (для сравнения: метанол-1, этен-1 и т.д.).

Программирование логического блока, осуществляющего синтаксический контроль и семантический анализ названий молекул, выглядит значительно проще, если к несистематическим фрагментам отнести также ФТОР, ХЛОР, НИТРО, АМИНО и т.д..

Если при анализе названия встречается несистематический фрагмент - программа обращается к библиотеке несистематических названий, содержащей их систематические аналоги, пытаясь отыскать там встреченный текст и необходимую для его интерпретации информацию.

Бензол — циклогексатриен-1,3,5

Пиридин — 1-аза-бензол.

7. Построение линейных атомных цепей

Результатом анализа названия является последовательность нециклических и циклических фрагментов, составляющих молекулу, методы построения которых достаточно хорошо разработаны [1]. В качестве примера приведем способ построения линейных цепочек.

Назовем стандартной систему координат (рис. 1), позволяющую вычислить координаты первых трех атомов по формулам:

$$\begin{array}{lll} x_1 = R_{12}cos\alpha, & x_2 = 0, & x_3 = R_{23} \\ y_1 = R_{12}sin\alpha, & y_2 = 0, & y_3 = 0 \\ z_1 = 0, & z_2 = 0, & z_3 = 0 \end{array} \qquad (1)$$

Легко, также, вычислить координаты четвертого атома:

$$\begin{aligned} x_4 &= x_3 - R_{34}cos\beta \\ y_4 &= R_{34}sin\beta cos\varphi \\ z_4 &= R_{34}sin\beta sin\varphi \end{aligned} \qquad (2)$$

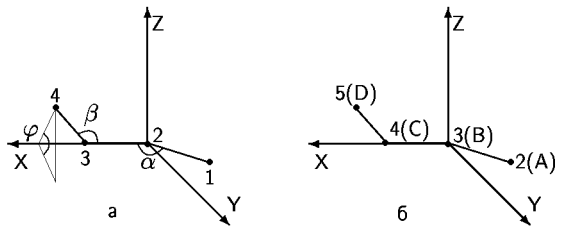

Рис. 1: Ориентация фрагмента молекулы в стандартной системе координат.

Теперь координаты всех атомов цепи могут быть вычислены с помощью следующего алгоритма:

1) Координаты первых четырех атомов вычисляются по формулам (1, 2).

2) Выбираются три опорных атома A, B, C с ранее найденными координатами (X, Y, Z).

3) Переносим начало координат в точку опорного атома B:

$$\Delta X_i = X_i - X_b, \Delta Y_i = Y_i - Y_b, \Delta Z_i = Z_i - Z_b$$

4) Вычисляем координаты атомов A, B, C и пристраиваемого атома D в стандартной системе координат по формулам (1, 2).

5) Полученные координаты связаны ортогональным преобразованием A

$$\Delta X = a_{11}x + a_{12}y + a_{13}z$$

$$\Delta Y = a_{21}x + a_{22}y + a_{23}z$$

$$\Delta Z = a_{31}x + a_{32}y + a_{33}z$$

элементы которого удается выразить следующим образом:

$$a_{11} = \Delta X_c/x_c,\ a_{12} = (\Delta X_a - a_{11}x_a)/y_a,$$

$$a_{21} = \Delta Y_c/x_c,\ a_{22} = (\Delta Y_a - a_{21}x_a)/y_a,$$

$$a_{31} = \Delta Z_c/x_c,\ a_{32} = (Z_a - a_{31}x_a)/y_a,$$

$$a_{13} = a_{21}a_{32} - a_{31}a_{22}$$

$$a_{23} = a_{31}a_{12} - a_{11}a_{32},$$

$$a_{33} = a_{11}a_{22} - a_{21}a_{12}$$

(особый случай $_a = R_{ab}\sin\alpha = 0$ возникает в производных ацетилена и легко исключается выбором в качестве атомов A, B и C другого, нелинейного фрагмента).

Лишь три из девяти матричных элементов a_{ij} независимы, справедливость связывающих их условий, накладываемых ортогональностью линейного преобразования A, может быть проверена непосредственно.

6) Координаты атома D (x_d, y_d, z_d) преобразуются в исходную систему координат:

$$\begin{pmatrix} X_d \\ Y_d \\ Z_d \end{pmatrix} = A \cdot \begin{pmatrix} x_d \\ y_d \\ z_d \end{pmatrix} + \begin{pmatrix} X_b \\ Y_b \\ Z_b \end{pmatrix}$$

и процесс повторяется с пункта 2 до полного построения модели.

8. Структура и области возможного применения комплекса

Создаваемый программный комплекс ГРИН (ГРафический Интерпретатор химических Названий) предполагает разработку следующих блоков:

1. Ввод, синтаксический и семантический анализ названия молекулы на естественном химическом языке. Для обработки рациональных и тривиальных названий создается база данных, содержащая соответствующие систематические аналоги.

2. Результатом анализа названия является последовательность нециклических и циклических фрагментов. Далее компьютерная модель молекулы строится путем последовательного вызова подпрограмм построения линейных и циклических цепей.

3. Необходимые для работы подпрограмм геометрические параметры молекул вызываются из специально созданных файлов в соответствии с типами образующих их атомов.

В настоящее время комплекс реализован для предельных линейных, разветвленных и циклических углеводородов. В дальнейшем его лексические возможности будут расширены на другие классы химических соединений. Это позволит эксплуатировать комплекс в следующих вариантах:

1. Справочник по химическому и геометрическому строению химических соединений (аналитических реактивов, природных соединений, фармацевтических препаратов и т.д.). В этом случае комплекс необходимо дополнить блоком оптимизации молекулярной геометрии, например, в рамках метода молекулярной механики. При этом оптимизацию геометрии и конформационный поиск можно будет организовать не в декартовых, а в естественных переменных, что значительно ускорит получение результатов.

2. Альтернативный способ ввода в ЭВМ информации о строении молекулы. При этом результатом работы комплекса будет файл, являющийся входным для программ, проводящих ее дальнейшую обработку.

3. Универсальный аналог геометрических подпрограмм, связывающих декартовы координаты атомов со структурными параметрами. Например, объединение комплекса с программой обработки данных электронографического эксперимента (KCED) позволит исключить медленную стадию "ручного"написания и отладки геометрической подпрограммы.

Литература.

1. Е.Г.Атавин, Вычисление декартовых координат атомов в больших молекулах, Ж.Общей химии, принято в печать 12.1998.

2. Mastryukov V. S., Simonsen S. H. // Molecular Structure Research. 1996. Vol. 2. P. 163-189.

3. H.Goto, A Revised Nomenclature for the Ring Conformation and a Note on the Conformational Distance in Cyclododecane, Tetrahedron Vol. 48, No. 35, pp.7131-7144, 1992.