

# ПРИЛОЖЕНИЯ ПОЛНЫХ ИНВАРИАНТОВ ГРАФОВ ДЛЯ СОСТАВЛЕНИЯ БАЗ ОРГАНИЧЕСКИХ МОЛЕКУЛ Кришталева А.Е.<sup>1</sup>, Байгунусов Р.С.<sup>2</sup>

<sup>1</sup>Кришталева Анна Евгеньевна – студент,  
кафедра прикладной математики;

<sup>2</sup>Байгунусов Руслан Сергеевич – студент,  
кафедра информатики и программного обеспечения вычислительных систем,  
Национальный исследовательский университет  
Московский институт электронной техники,  
г. Москва

**Аннотация:** рассматривается линейная нотация – полный инвариант графа, который можно использовать для альтернативного описания конечных графов, в частности, для составления баз данных органических молекул, приводятся преимущества линейной нотации перед одной уже существующей системой правил описания структуры вещества.

**Ключевые слова:** изоморфизм графов, линейная нотация, SMILES.

В работе [1] была введена линейная нотация абстрактного графа  $I[G]$ . Было продемонстрировано, что она является полным инвариантом графа и по всем свойствам во многом аналогична обычным графам  $G$ . Факт того, что  $I[G]$  является полным инвариантом, может послужить основанием использования нотации  $I[G]$  в качестве альтернативного способа хранения графов. В частности, для составления баз органических молекул.

Для этих целей уже используется SMILES (Simplified Molecular Input Line Entry Specification) — система правил состава и структуры молекулы химического вещества с использованием строки символов ASCII. Первоначальный ее вариант был разработан Дэвидом Вейнингером в конце 1980-х. Другими словами, SMILES – это строка, полученная путем перечисления атомных символов и типов связей, использующая обход в глубину дерева молекулярного графа, где вершины представляют атомы, а ребра – химические связи. И здесь возникла проблема: не существует однозначного порядка обхода вершин в молекулярном графе, так как обход в глубину из разных вершин будет давать разные результаты. Чтобы преодолеть этот недостаток, создатели SMILES разработали двухэтапный алгоритм под названием CANGEN, который позволял создавать уникальную строку SMILES для каждого молекулярного соединения.

Алгоритм составления уникальной SMILES состоит из следующих двух этапов:

1. Первый этап, CANON, включает канонизацию структуры, представленной в виде молекулярного графа. На этом этапе молекулярный граф помечается каноническими метками. Каждому атому (вершине) присваивается числовая метка на основе ее топологии.

2. Второй этап, GENES, генерирует уникальное обозначение SMILES, представляющее собой обход молекулярного графа в глубину. GENES выбирает исходный атом и принимает решения ветвления, ссылаясь на канонические метки по мере необходимости.

Подробный алгоритм составления строки SMILES описан в [2].

Для большинства химических структур алгоритм CANGEN, генерирует действительно уникальные записи. Однако есть исключения. И причина их возникновения довольно проста: если граф достаточно симметричен для этапа CANON, алгоритм SMILES можно генерировать различные канонические метки для вершин молекулярного графа

Рассмотрим результат работы алгоритма построения SMILES на примере 3,5-диэтилтолуола. Он хранится в базе данных NCI с NSC-номером 62141.

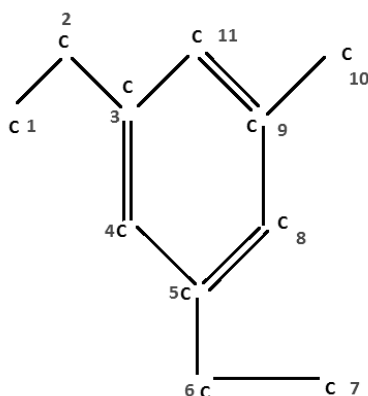


Рис. 1. Молекулярный граф 3,5-диэтилтолуола

После этапа CANON имеем две альтернативные разметки атомов, зависящие от выбора начальной вершины, согласно приоритетам

Таблица 1. Канонический порядок вершин, соответствующий обходу в глубину из узла 1

| Номер вершины      | 1 | 2 | 3  | 4 | 5  | 6 | 7 | 8 | 9 | 10 | 11 |
|--------------------|---|---|----|---|----|---|---|---|---|----|----|
| Каноническая метка | 1 | 4 | 10 | 8 | 11 | 5 | 2 | 7 | 9 | 3  | 6  |

Таблица 2. Канонический порядок вершин, соответствующий обходу в глубину из узла 7

| Номер вершины      | 1 | 2 | 3  | 4 | 5  | 6 | 7 | 8 | 9 | 10 | 11 |
|--------------------|---|---|----|---|----|---|---|---|---|----|----|
| Каноническая метка | 2 | 4 | 10 | 8 | 11 | 5 | 1 | 7 | 9 | 3  | 6  |

После этапа GENES получаем два варианта строки SMILES:

CCC1 = CC(= CC (= C1)C)CC

с началом в вершине «1». И

CCC1 = CC(= CC (= C1)CC)C

с началом в вершине «7».

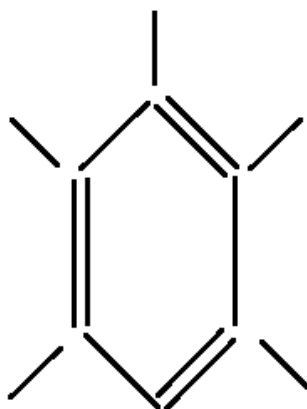


Рис. 2. Молекулярный граф пентаметилбензола

Или вот две различные строки SMILES для пентаметилбензола с NSC ID 1889.

CC1 = C(C)C(= C(C)C(= C1)C)C

CC1 = CC(= C(C)C(= C1(C))C)C

Таким образом, получается, что алгоритм CANGEN составления SMILES не гарантирует уникальность каждой записи. Для решения этой проблемы можно использовать линейную нотацию  $I[G]$ , предложенную Назаровым М. Н. и подробно описанную в [1]. Ниже приведем ее краткое описание, реализацию построения для молекулярных графов и преимущества перед SMILES.

Линейная нотация  $I[G]$  для связного графа – это специальная строка символов, позволяющая уникальным образом описать графовую структуру.

Алгоритм составления линейной нотации включает в себя следующие этапы:

Этап 1. Нахождение последовательностей вершин, соответствующей макси-коду графа.

Этап 2. Определение классов автоморфизма вершин.

Этап 3. Определение классов автоморфизма ребер.

Этап 4. Составление линейной нотации  $I[G]$ . Подробный алгоритм данного этапа описан в [1].

Использование линейной нотации для органических молекул позволит компактно хранить большие структурные соединения, при этом гарантируется уникальность каждой записи.

Рассмотрим результат работы алгоритма построения линейной нотации на примере 3,5-диэтилтолуола.

После этапа 1 получаем две последовательности вершин, соответствующих макси-коду молекулярного графа:

5 4 8 6 3 9 7 11 2 10 1  
3 4 11 2 5 9 1 8 6 10 7

На следующем этапе определяем классы автоморфизма вершин

Таблица 3. Классы автоморфизма вершин молекулярного графа 3,5-диэтилтолуола

| Номер вершины      | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|--------------------|---|---|---|---|---|---|---|---|---|----|----|
| Класс автоморфизма | 6 | 4 | 1 | 2 | 1 | 4 | 6 | 3 | 5 | 7  | 3  |

На этапе определения классов автоморфизма ребер получаем

Таблица 4. Классы автоморфизма ребер молекулярного графа 3,5-диэтилтолуола

| Ребро | (1,2) | (2,3) | (3,4) | (3,11) | (4,5) | (5,6) | (5,8) | (6,7) | (8,9) | (9,10) | (9,11) |
|-------|-------|-------|-------|--------|-------|-------|-------|-------|-------|--------|--------|
| Класс | 5     | 3     | 1     | 2      | 1     | 3     | 2     | 5     | 4     | 6      | 4      |

После этапа 4 получаем линейную нотацию  $I[G]$  и молекулярный граф с метками:

1[(1) – 2[(1) – 1[(2) – 3[(4) – 5[(4) – 3[(2) – #1](6) – 7[]]](3) – 4[(5) – 6[]]](3) – 4[(5) – 6[]]]

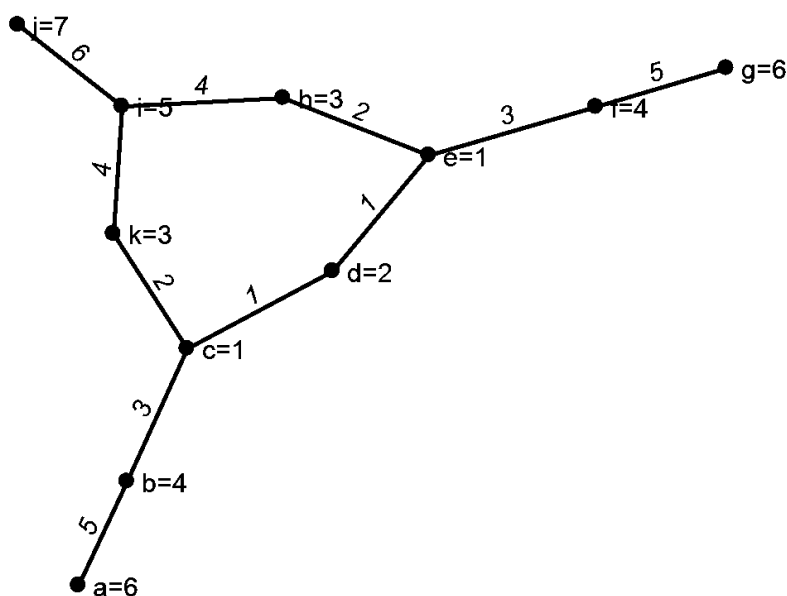


Рис. 3. Молекулярный граф 3,5-диэтилтолуола с метками

Для молекулярного графа пентаметилбензола, структурная формула которого представлена на рисунке 2, линейная нотация имеет вид:

1[(1) – 1[(1) – 1[(2) – 2[(5) – 4[]](6) – 4[(6) – 2[(2) – #1; (5) – 4[]]](3) – 3[]](4) – 3[]](3) – 3[]]

А сам граф с метками выглядит следующим образом:

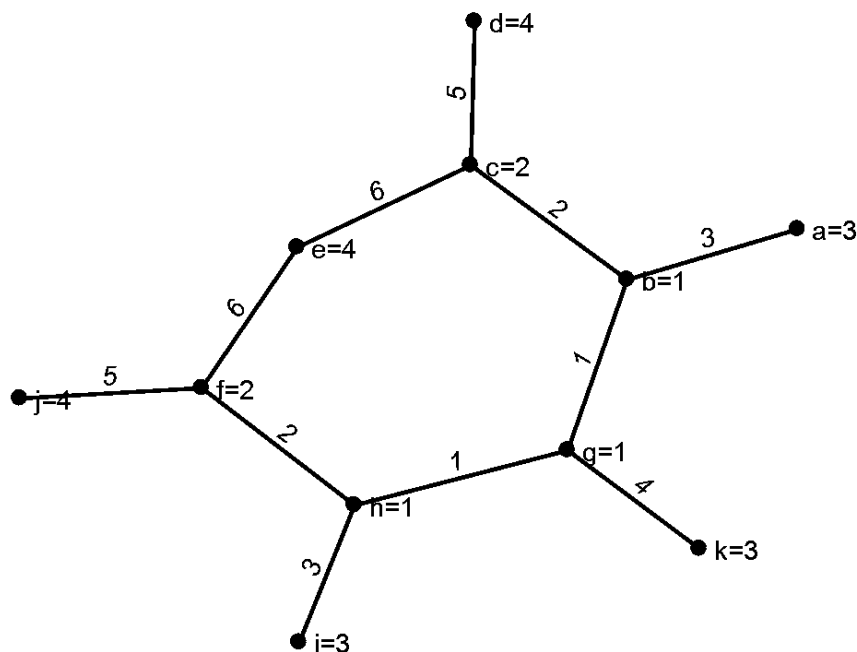


Рис. 4. Молекулярный граф пентаметилбензола с метками

Процесс составления линейной нотации  $I[G]$  достаточно трудоемкий. Нахождение макси-кода графа, поиск классов автоморфизма вершин и ребер требуют больших вычислительных затрат. Тем не менее, при описании молекулярного графа таким образом получаем следующие преимущества:

- На линейной нотации наглядно видны все симметрии графа через индексы классов автоморфизма вершин и ребер.
- Проверка на изоморфизм двух графов имеет сложность  $O(n^2)$ .

Самым важным преимуществом линейной нотации перед SMILES является то, что линейная нотация гарантирует уникальность записи любого химического соединения. Поэтому ее можно взять за основу для составления баз данных органических молекул.

#### Список литературы

1. Назаров М.Н. Об альтернативном способе задания конечных графов // Прикладная дискретная математика, 2015. № 3. С. 83-94.
2. Weininger D., Weininger A. and Weininger J. SMILES. 2. Algorithm for generation of unique SMILES notation // J. Chem. Inf. Comput. Sci., 1989. V. 29. № 2. P. 97–101.
3. Weininger David. SMILES, a Chemical Language and Information System 1: Introduction to Methodology and Encoding Rules, Medicinal Chemistry Project. Pomona College, 1988.
4. Зыков А.А. Основы теории графов. М.: Вузовская книга, 2004. 664 с.