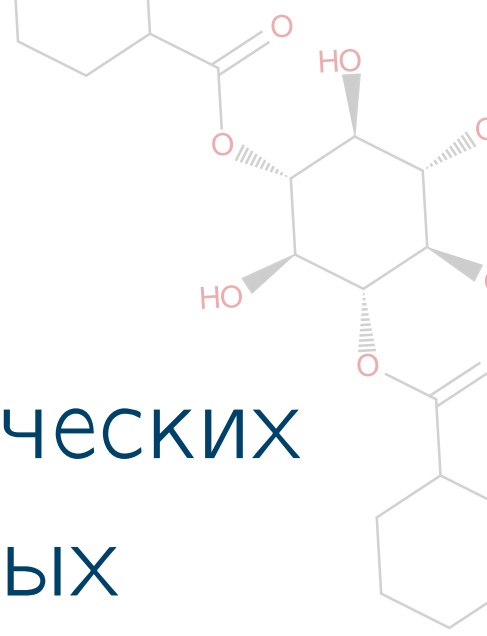
GGA SOFTWARE SERVICES Open-source software



Подструктурный поиск химических соединений в базах данных

Михаил Рыбалкин

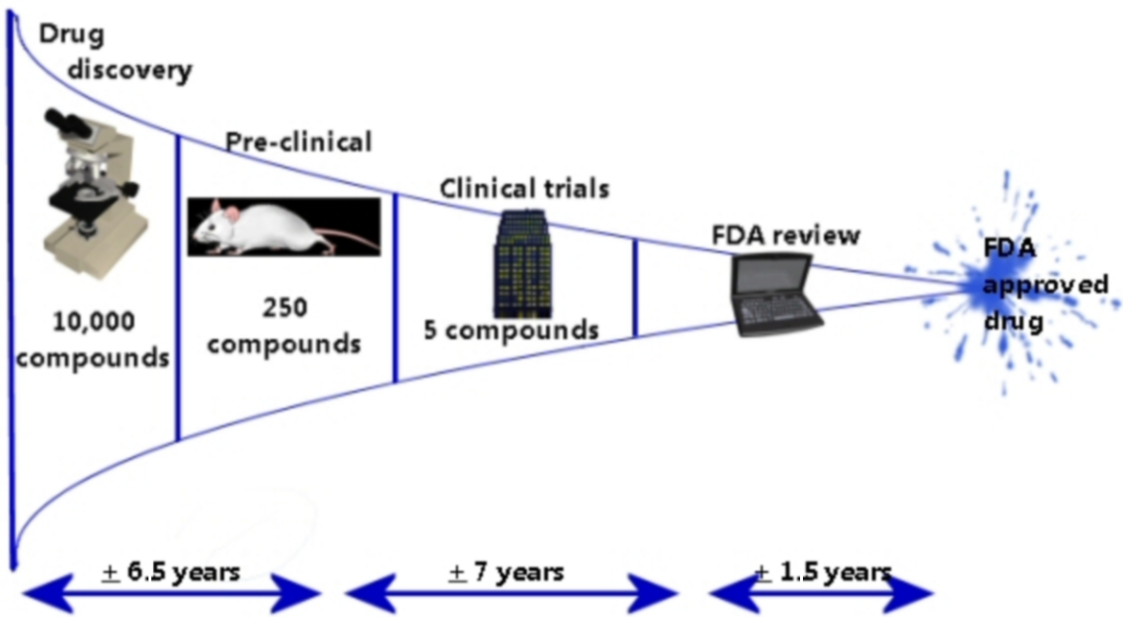
01.05.2011

Введение. Разработка лекарств

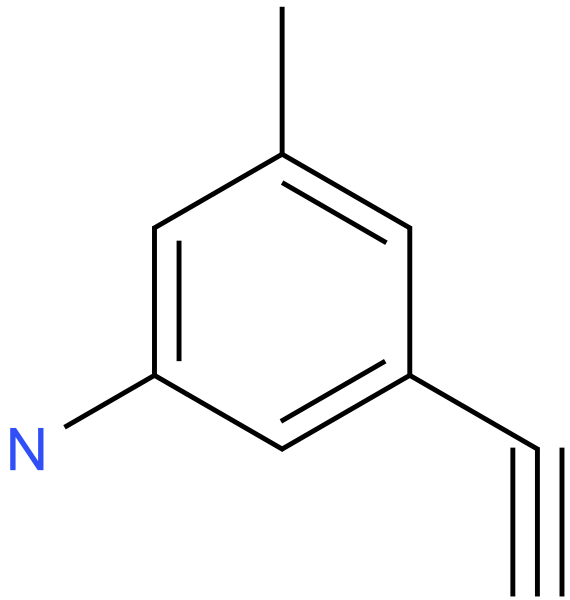
- Количество веществ $\sim$ 30-40 млн.
- Количество лекарственных веществ $\sim$ 1-2 тыс.
  Аналоги лекарств

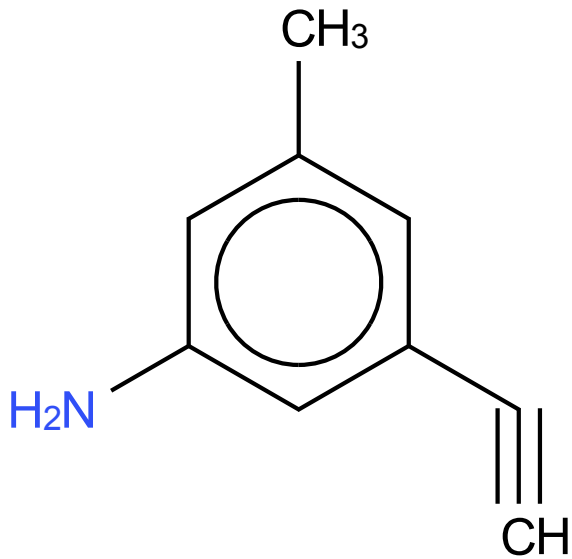
Введение. Разработка лекарств

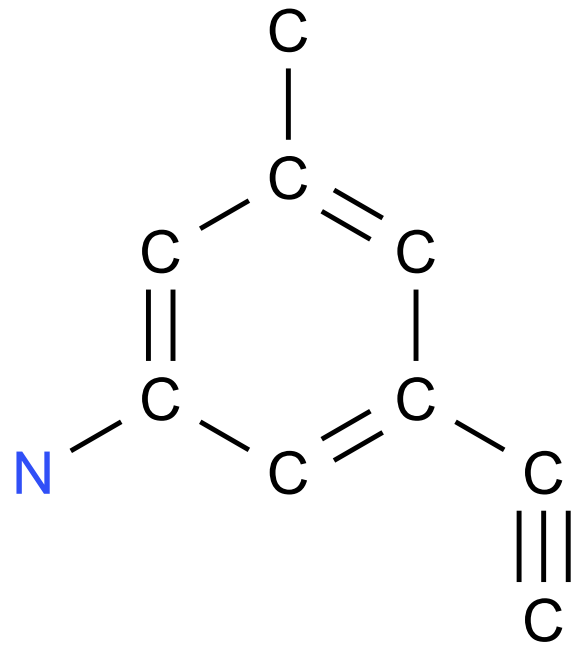
- Количество веществ $\sim$ 30-40 млн.
- Количество лекарственных веществ $\sim$ 1-2 тыс. Аналоги лекарств
- Разработка лекарств $\sim$ 10-15 лет



Отступление: химические обозначения



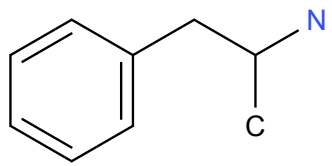




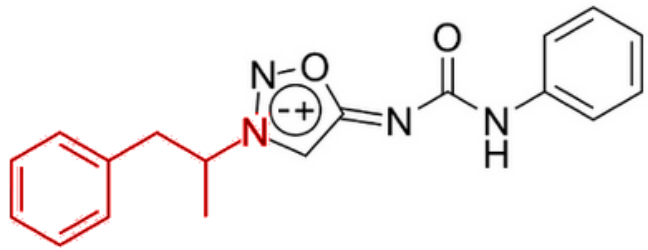
Свойства веществ

Амфетамин

Стимулятор центральной нервной системы



Повышение артериального давления

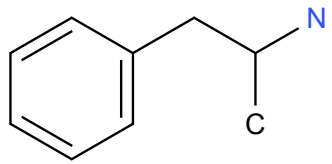


Мезокарб

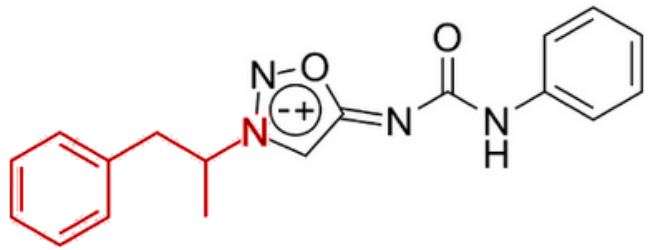
Свойства веществ

Амфетамин

Стимулятор центральной нервной системы

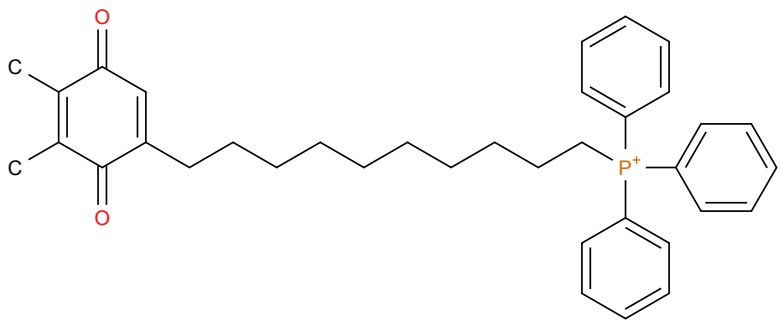


Повышение артериального давления



Мезокарб

Антиоксиданты

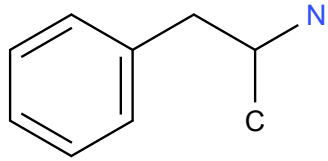


SkQ1

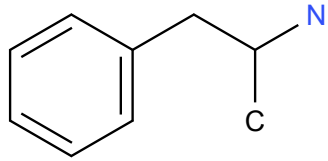
- Проникает в клетку
- Антиоксидант

Академик Скулачев В.П.

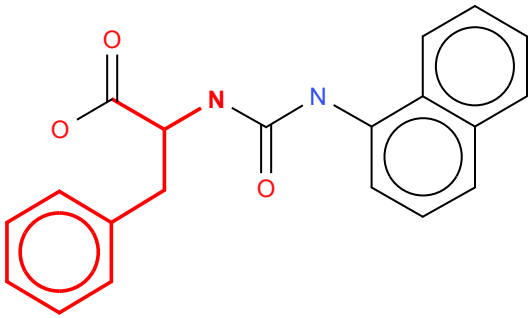
Виды поиска

Запрос: 

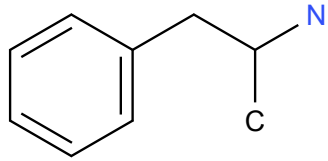
Виды поиска

Запрос: 

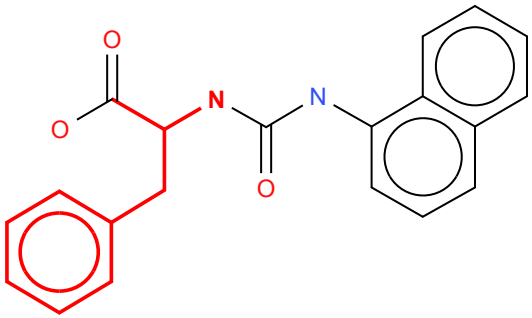
- Substructure (Подструктурный поиск)



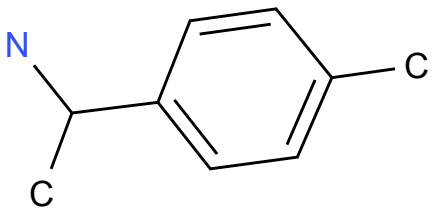
Виды поиска

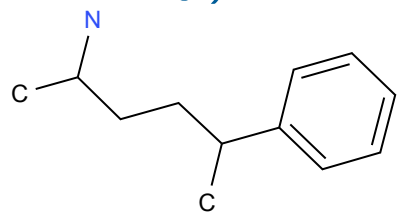
Запрос: 

- Substructure (Подструктурный поиск)

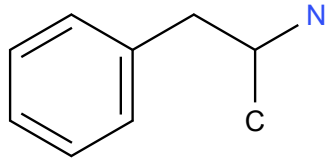


- Similarity (Поиск по сходству)

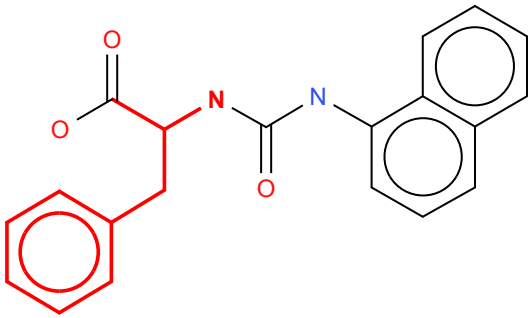




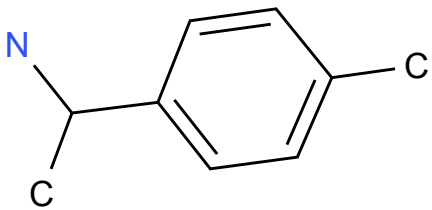
Виды поиска

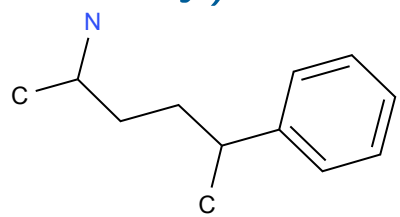
Запрос: 

- Substructure (Подструктурный поиск)

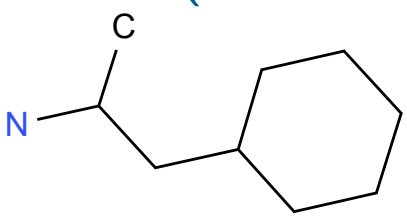


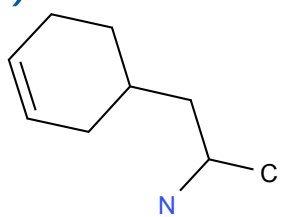
- Similarity (Поиск по сходству)





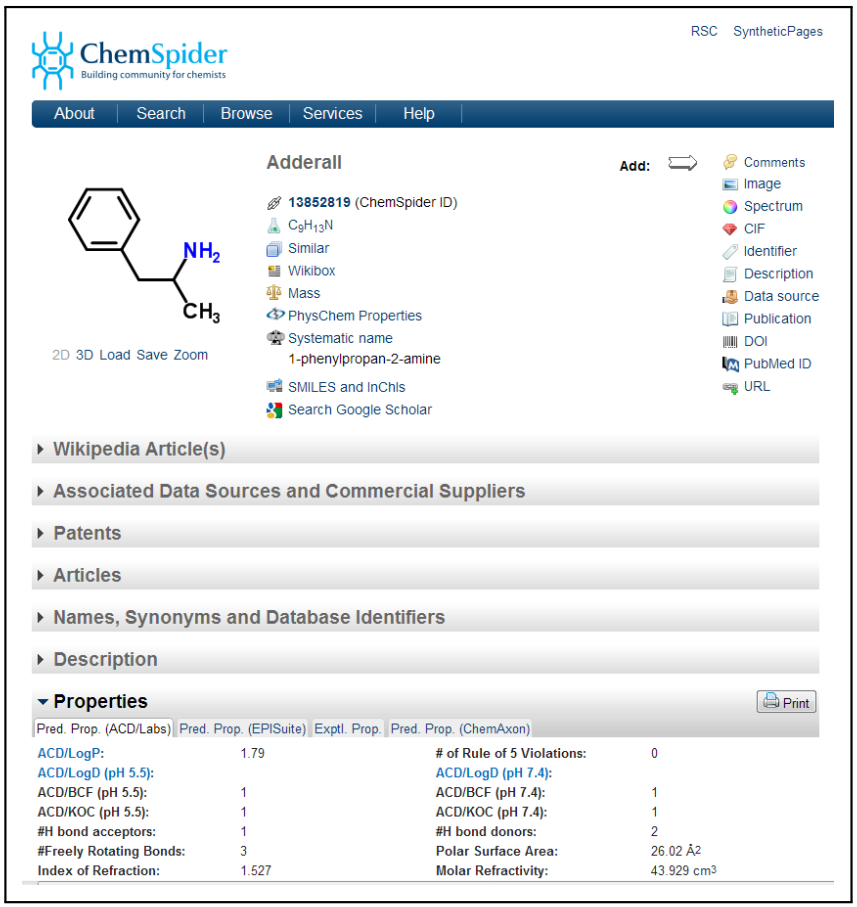
- Exact (Точный поиск)

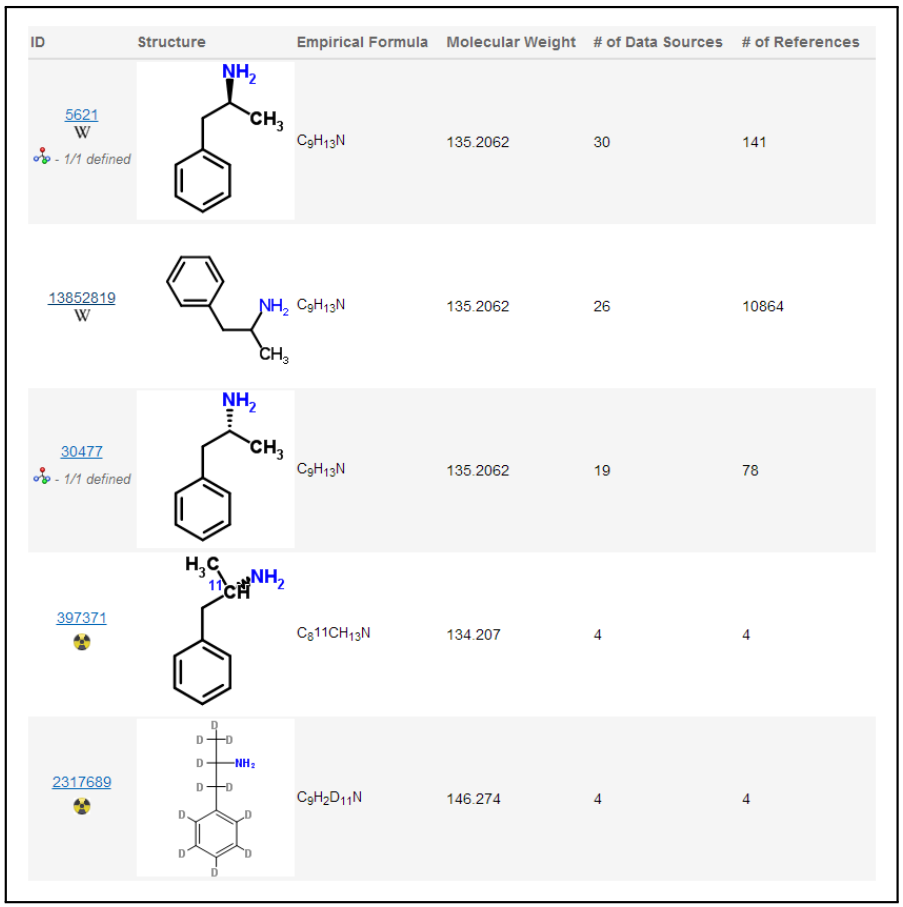




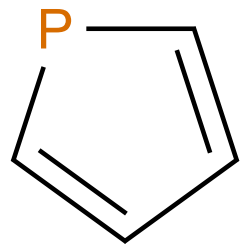
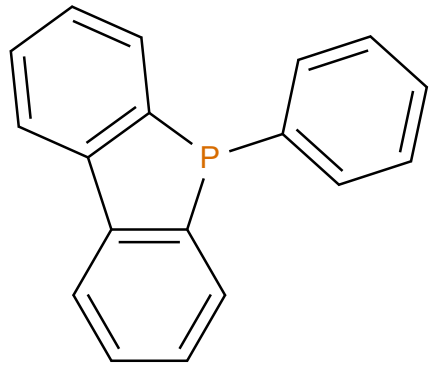
"Поисковики" химических соединений

- PubChem - 32 млн. молекул
- ChemSpider - 25 млн. молекул (использует Bingo)

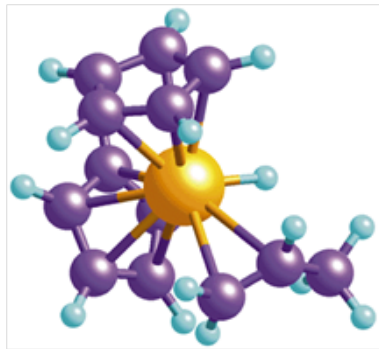
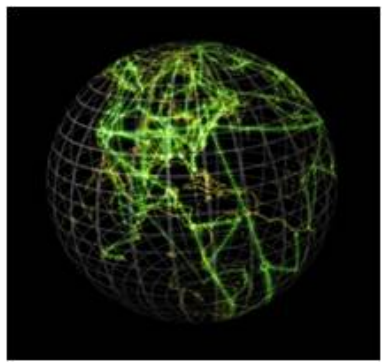




Структурированность данных

Текст	Графы	Изображения
Погодва		
▸ **Погода** в Санкт-Петербурге ▸ Подробный прогноз **погоды**		

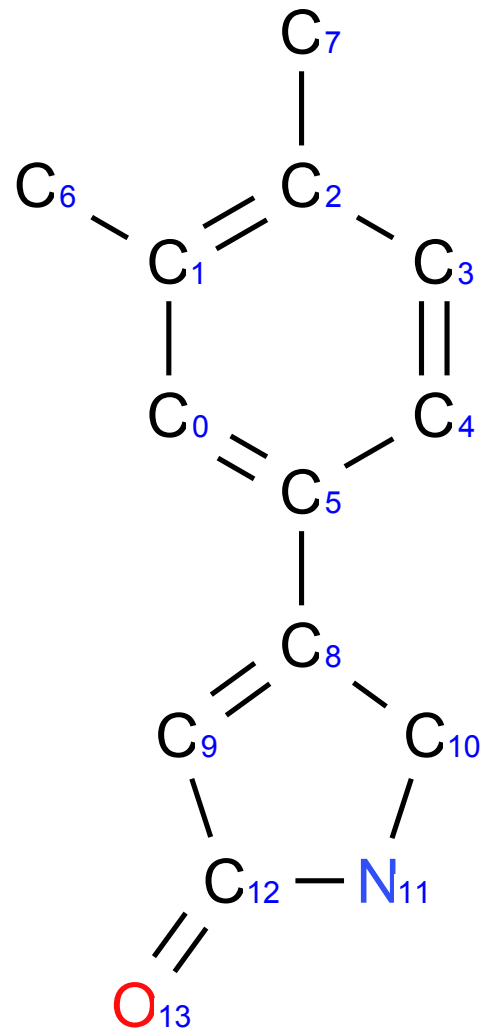
Использование графов

Химические соединения	
Схемы взаимодействия протеинов	
Социальные сети	

Исходные данные

Молекула - помеченный граф $G(V, E, f_V, f_E)$

- $|V| < 200$
- Почти все графы планарны
- $d(G) \leq 7 \quad (d(G) \leq 4)$
- $|f_V| < 118$ – атомы.
  Но есть дополнительные свойства: заряд, ...
- $|f_E| = 4$ – типы связей



Поиск во множестве молекул

GGA SOFTWARE SERVICES

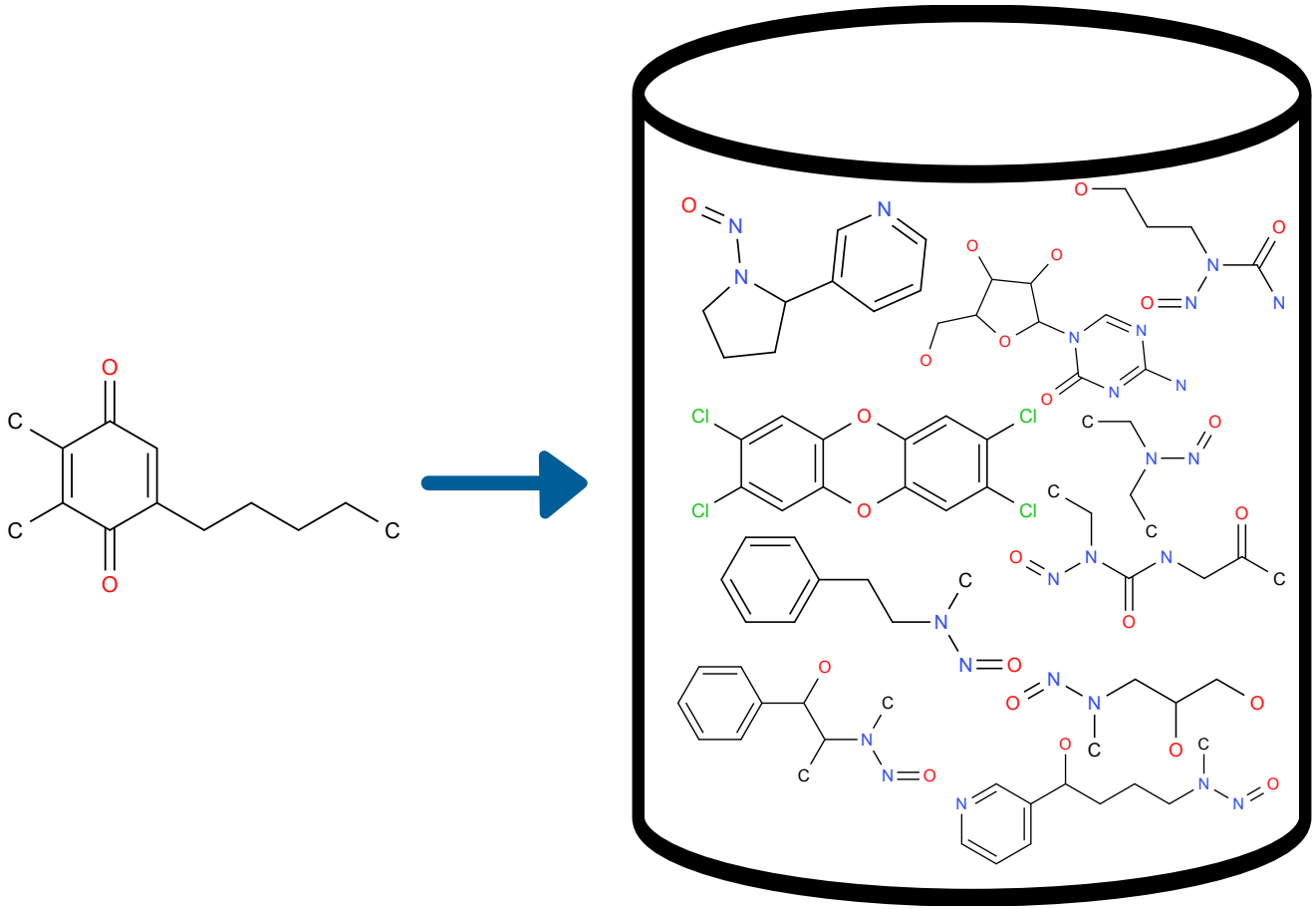
Дано:

- $D = \{G_1, G_2, G_3, ..., G_k\}$ – множество графов
- Q – граф-запрос ($|Q| < 20$)

Подструктурный поиск:

- Найти все граф, содержащие заданный подграф $D_Q = \{G \in D | Q \subset G\}$

Поиск во множестве молекул (2)



- Количество молекул $\sim$ 30 млн.
- Необходимо выбирать подмножество для поиска

Схема алгоритма

Преобработка данных – построение индекса:

- Сохранение дополнительной информации
  Скорость предобработки

Схема алгоритма поиска:

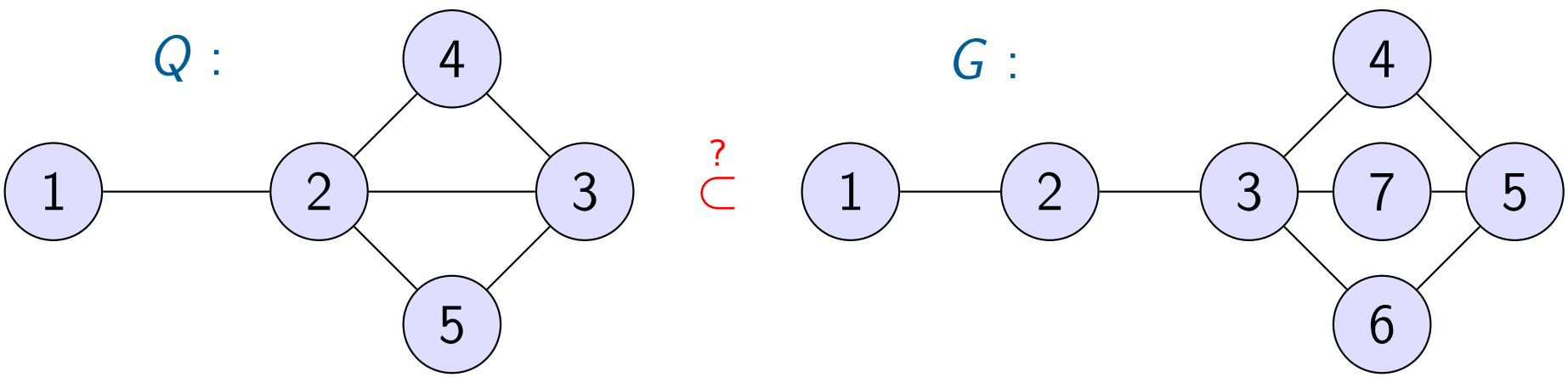
1. Выбор множества кандидатов
   Эффективность индекса
2. Поиск изоморфного подграфа для каждого кандидата
   Эффективность поиска изоморфного подграфа

Поиск изоморфного подграфа

Subgraph isomorphism problem - NP-полная задача
(Полиномиальный алгоритм?)

[Cordella 2004] – Поиск изоморфного подграфа для разреженных (ориентированных) графов

Рекурсивный обход с отсечением



Поиск изоморфного подграфа (2)

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Поиск изоморфного подграфа (2)

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

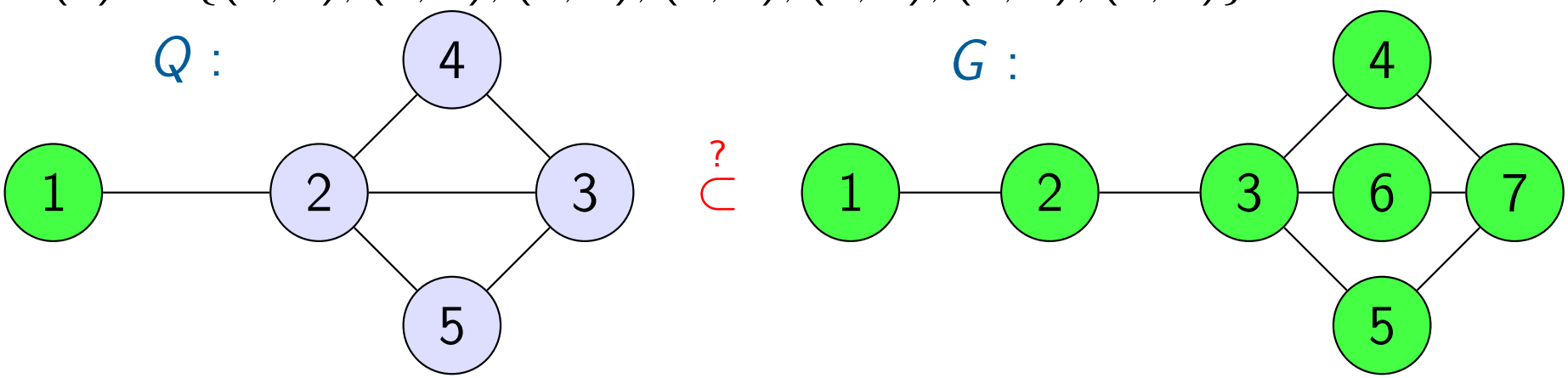
$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Шаг 1

$s = \{\}$

$T(s) = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(1,7)\}$



Поиск изоморфного подграфа (2)

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

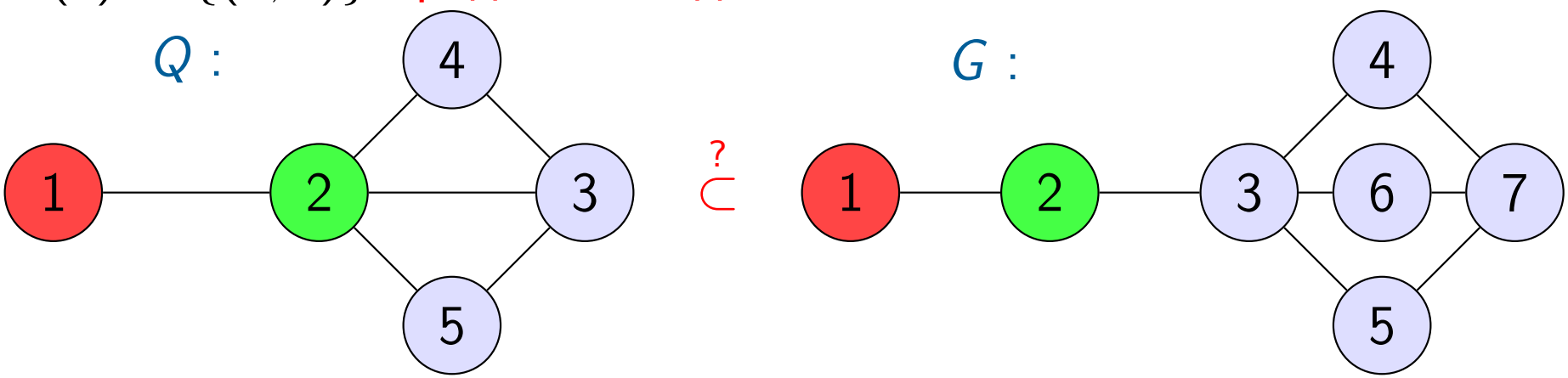
$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Шаг 2

$s = \{(1, 1)\}$

$T(s) = \{(2, 2)\}$ продолжить дальше нельзя



Поиск изоморфного подграфа (2)

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

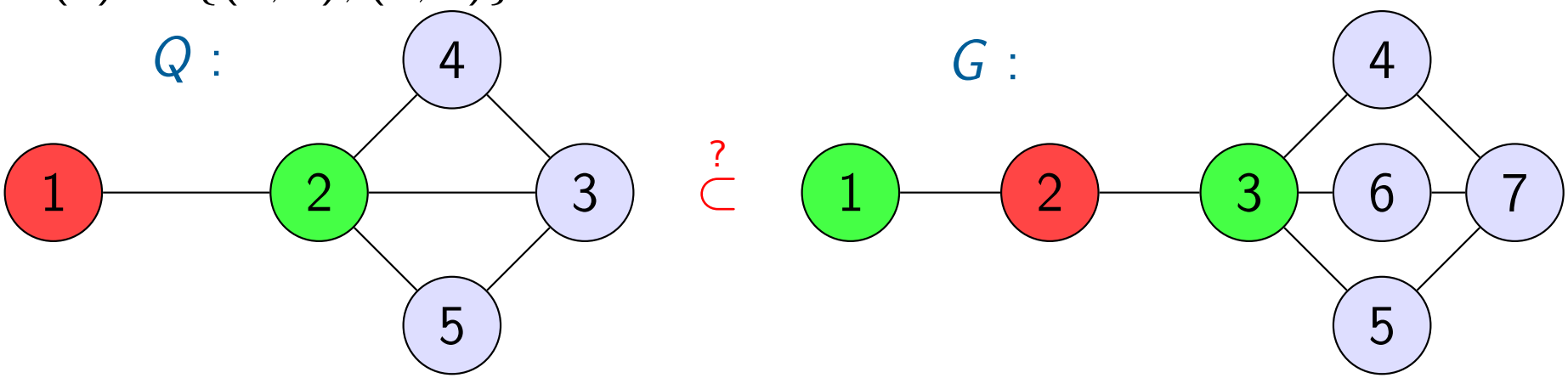
$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Шаг 3

$s = \{(1, 2)\}$

$T(s) = \{(2, 1), (2, 3)\}$



Поиск изоморфного подграфа (2)

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

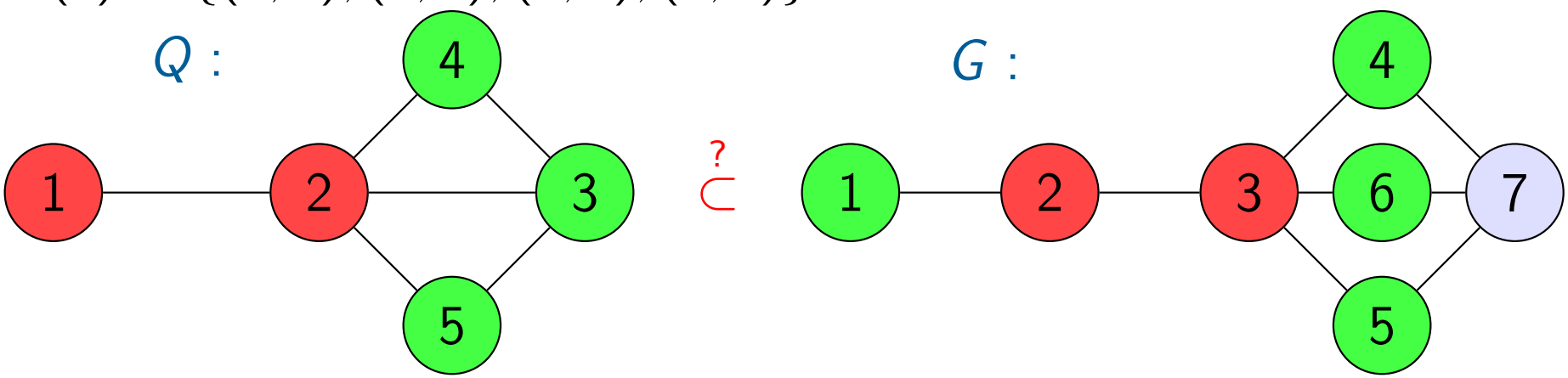
$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Шаг 4

$s = \{(1,2),(2,3)\}$

$T(s) = \{(3,1),(3,4),(3,5),(3,6)\}$



Поиск изоморфного подграфа (2)

GGA SOFTWARE SERVICES

Требуется проверить $Q \subset G$

$s \subset V(G) \times V(Q)$ – частичное отображение, состояние

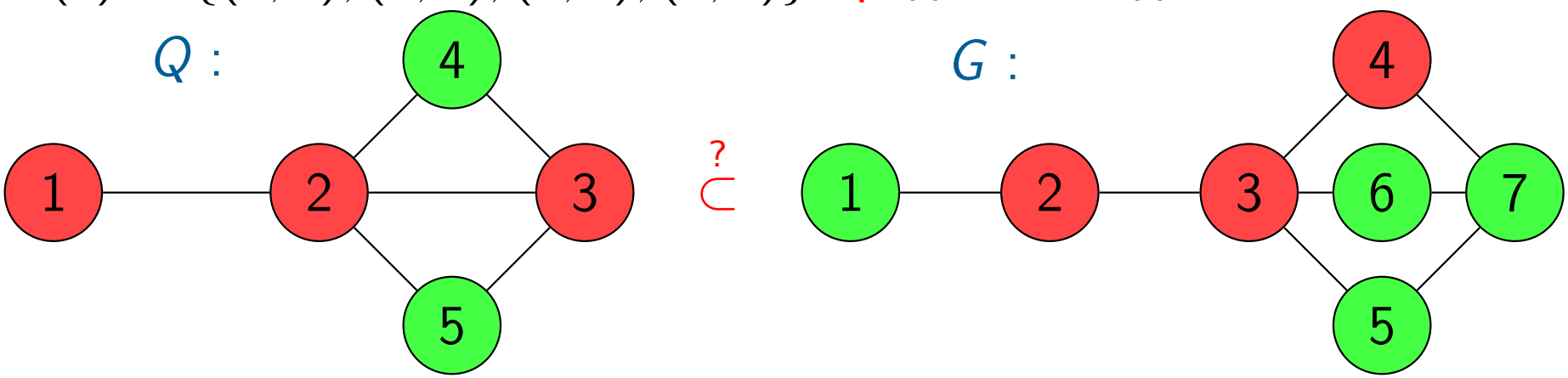
$T(s)$ – множество пар, которые можно добавить к s

$T(s) = \{\min V_s(Q)\} \times V_s(G)$

Шаг 5

$s = \{(1,2),(2,3),(3,4)\}$

$T(s) = \{(4,1),(4,5),(4,6),(4,7)\}$ продолжить дальше нельзя

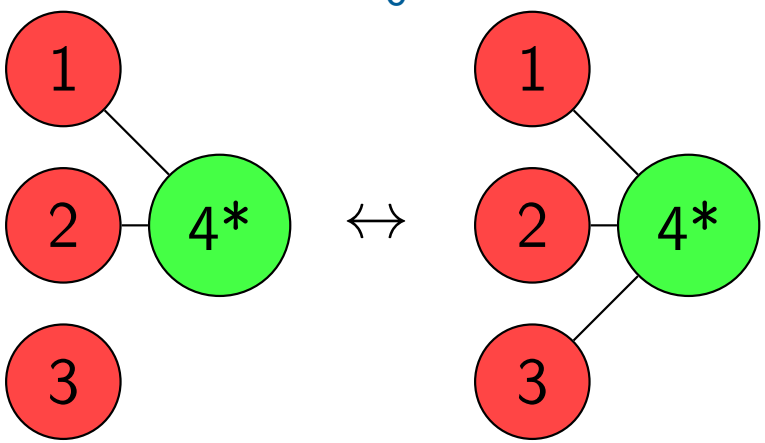


Поиск изоморфного подграфа (3)

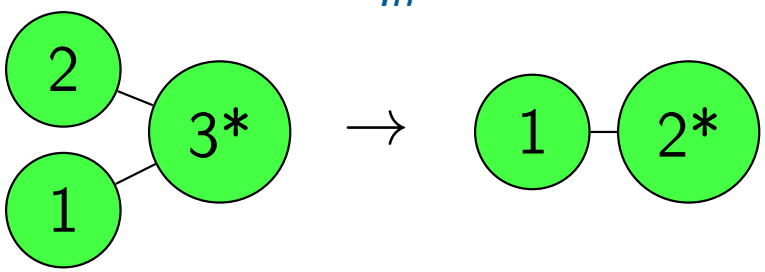
$F(s, q, g)$ – можно ли к s добавить (q, g) *(feasibility)*

Отсечение ветвей рекурсии:

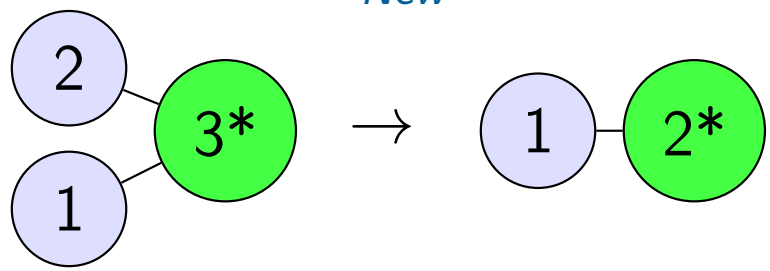
F_0



F_{In}

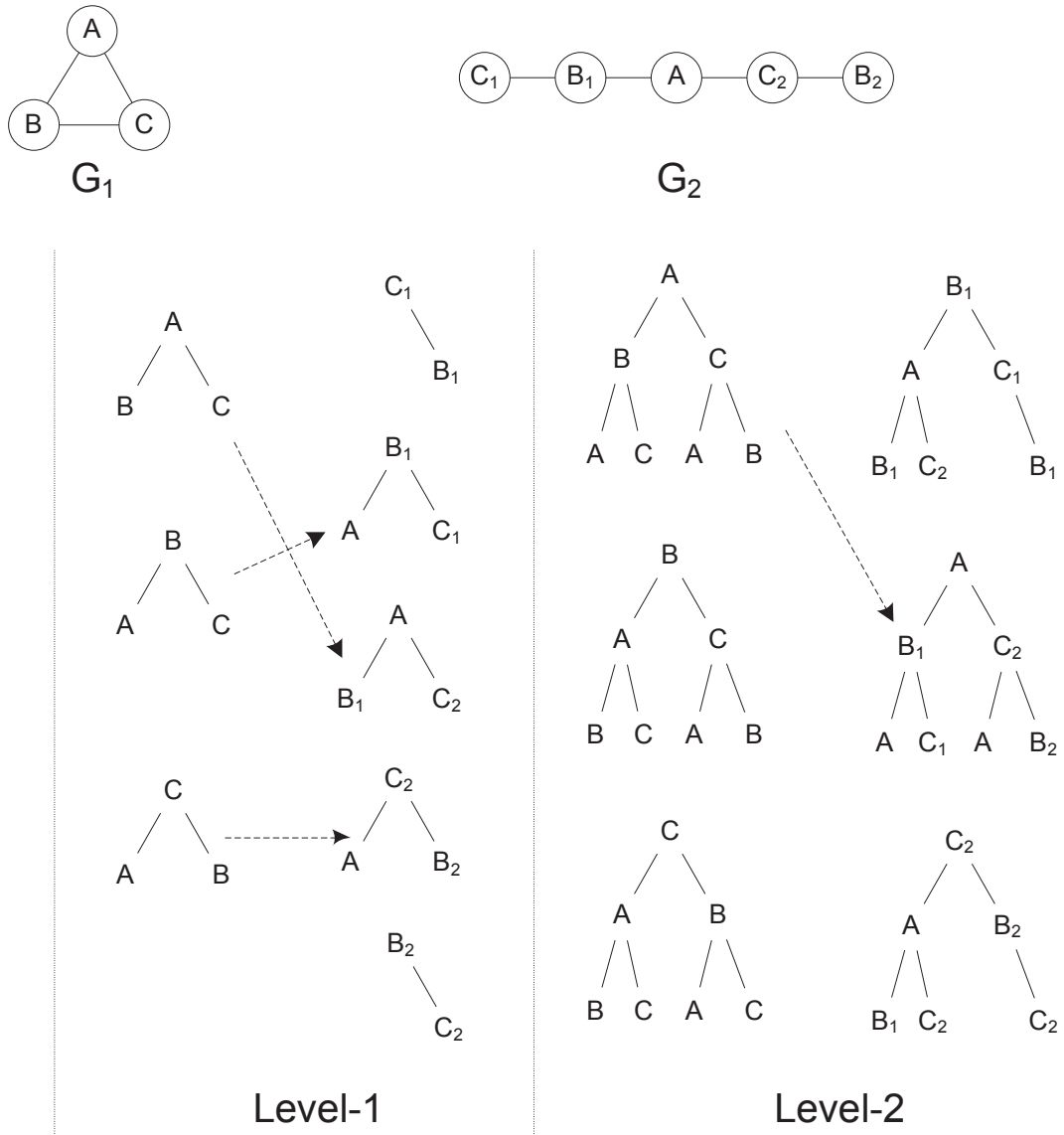


F_{New}



"Приближенный" поиск

[He 2006]: Closure-Tree: An Index Structure for Graph Queries

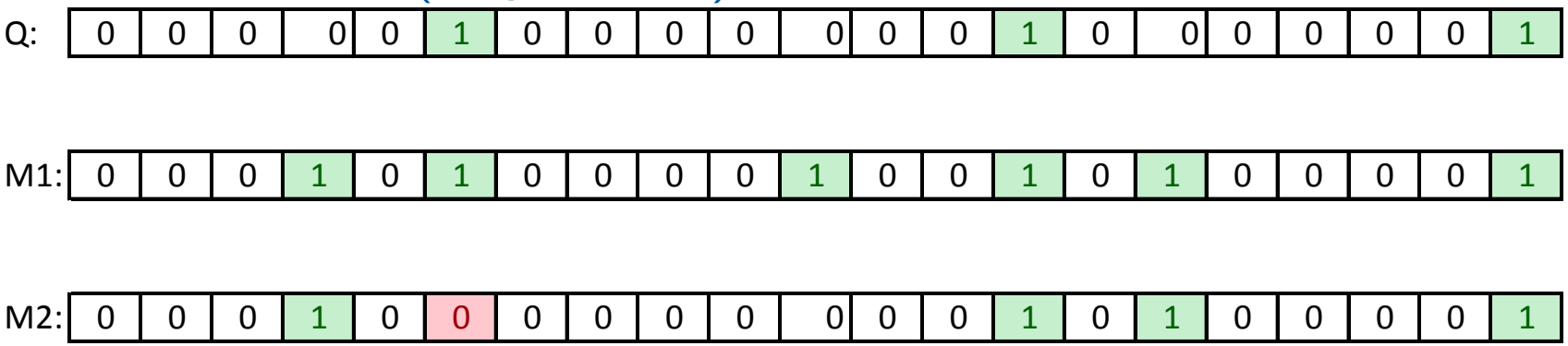


Возможные улучшения

- Инварианты окрестностей атомов
- Симметричность графов

Битовые отпечатки молекул

Фингерпринты (fingerprints)



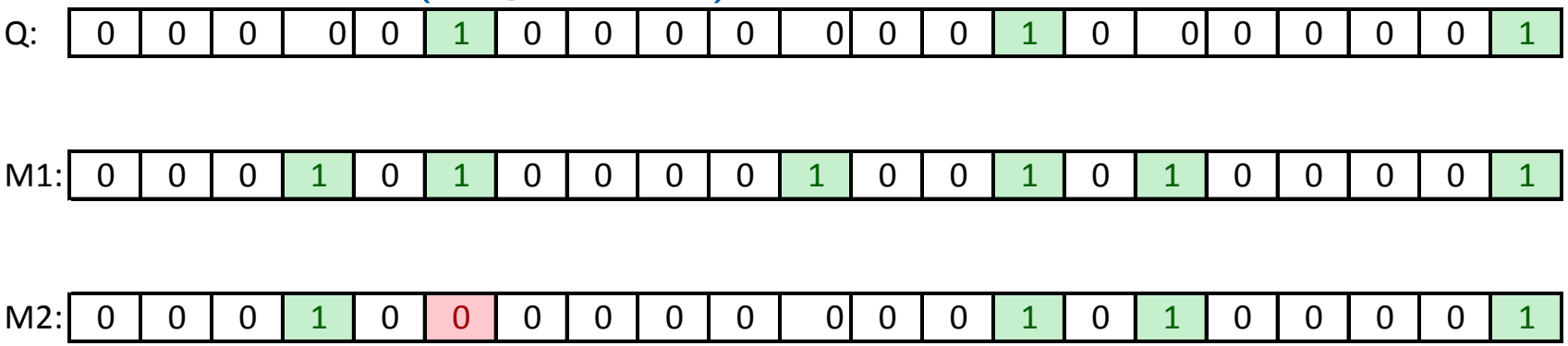
- Каждый бит соответствует свойству

Свойства:

- Фиксированный набор свойств

Битовые отпечатки молекул

Фингерпринты (fingerprints)



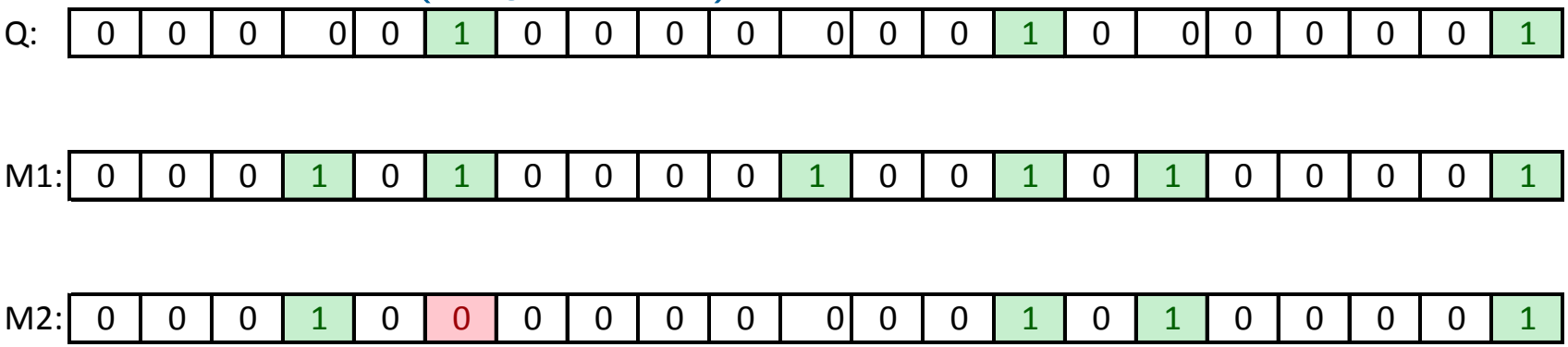
- Каждый бит соответствует свойству

Свойства:

- Фиксированный набор свойств
- Универсальные свойства – множество подграфов ограниченного размера

Битовые отпечатки молекул

Фингерпринты (fingerprints)



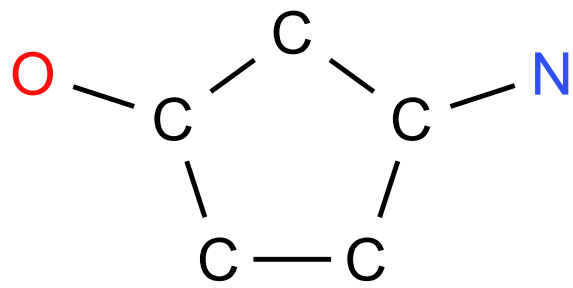
- Каждый бит соответствует свойству

Свойства:

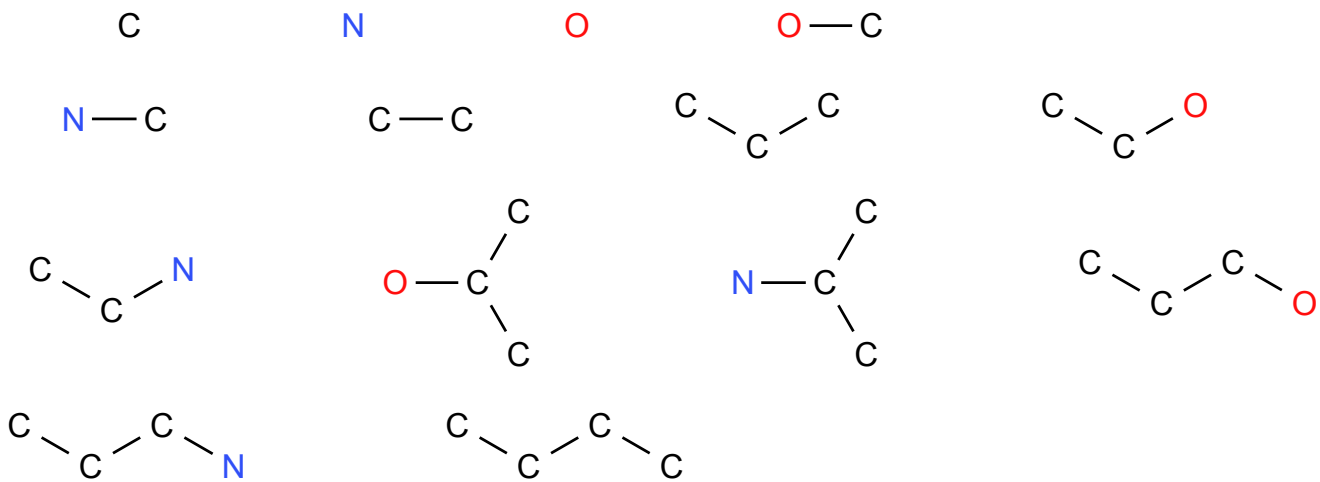
- Фиксированный набор свойств
- Универсальные свойства – множество подграфов ограниченного размера
  - Цепочки
  - Подграфы
  - Поддеревья
  - Циклы

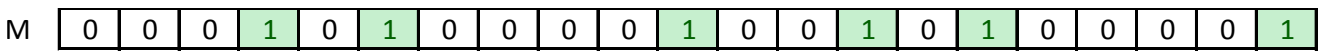
Построение битовых отпечатков

Граф:



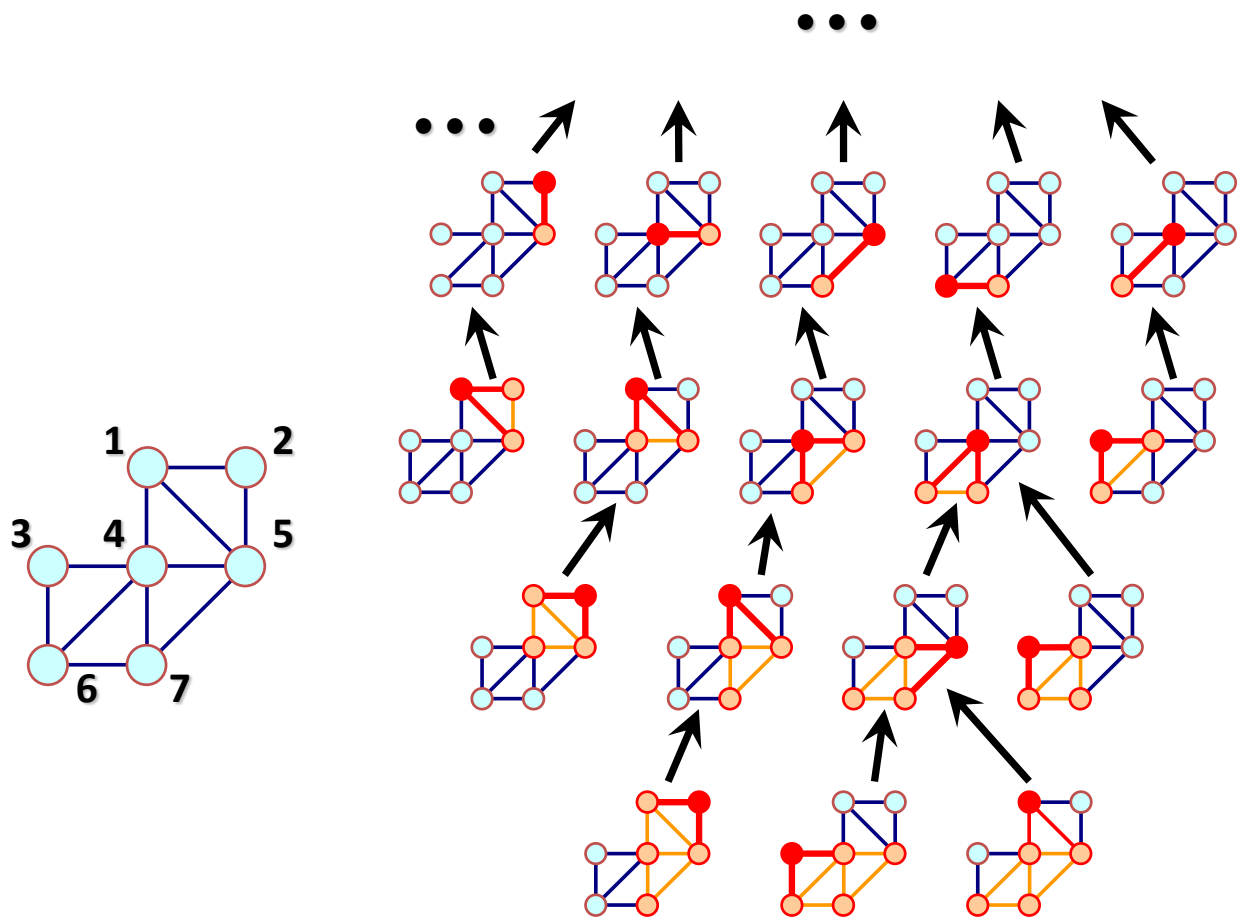
Поддеревья до 4 вершин:





Реверсивный поиск

[Avis 1993]: Reverse Search for Enumeration



Приложения: перебор подграфов, триангуляций, поддеревьев, вершин многогранника, ...

Организация индекса

Битовые отпечатки считываются с диска

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Q:	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
M1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
M3	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
M4	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0
M5	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M6	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
M7	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
M8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
M9	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
M10	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Организация индекса

Битовые отпечатки считываются с диска

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Q:	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
M1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
M3	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
M4	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0
M5	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
M6	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
M7	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
M8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
M9	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
M10	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

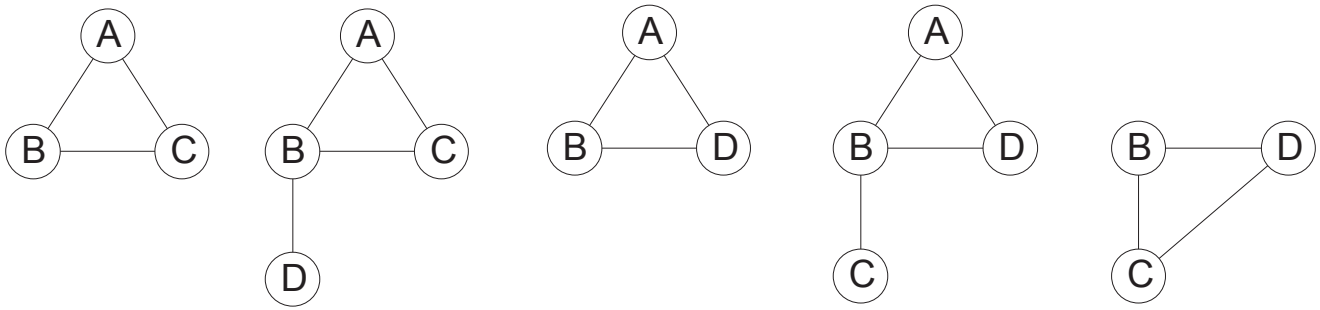
Транспонированное представление:

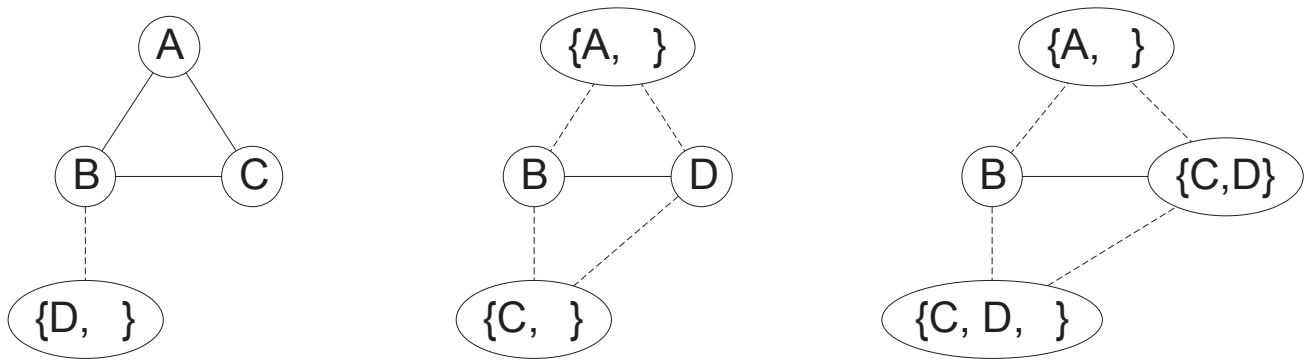
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
bit6:	0	0	0	1	0	0	1	0	0	0
bit14:	0	0	1	1	1	0	0	0	1	0
bit21:	0	0	1	1	0	0	1	0	0	0
bit23:	0	1	1	1	0	1	0	0	0	0

	0	0	0	1	0	0	0	0	0	0

Closure tree

[He 2006]: Closure-Tree: An Index Structure for Graph Queries





Особенности химического поиска

GGA SOFTWARE SERVICES

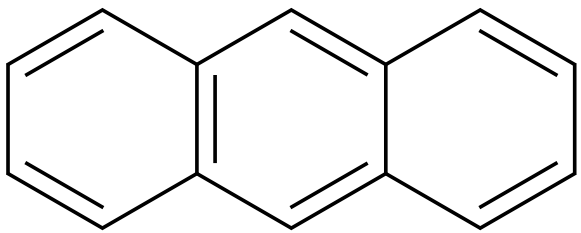
Химический поиск сложнее поиска подграфа

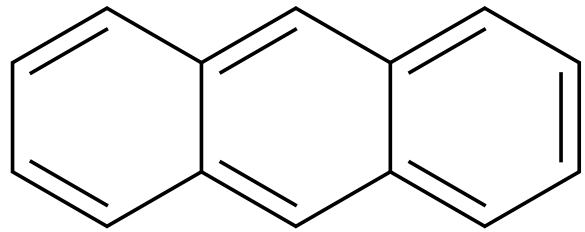
- Логические выражения
  SMARTS – SMiles ARbitrary Target Specification

Особенности химического поиска

Химический поиск сложнее поиска подграфа

- Логические выражения
  SMARTS – SMiles ARbitrary Target Specification
- Ароматичность

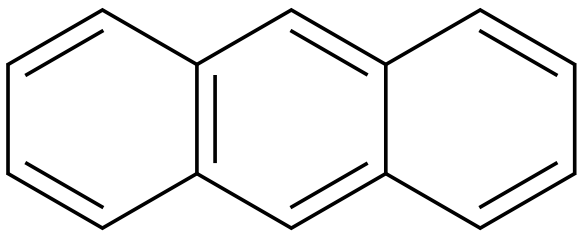


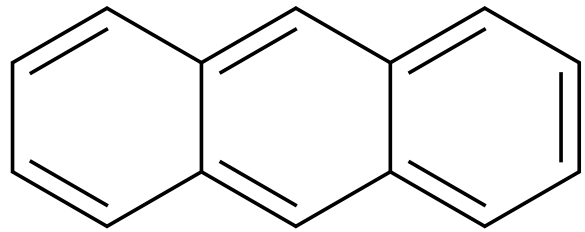


Особенности химического поиска

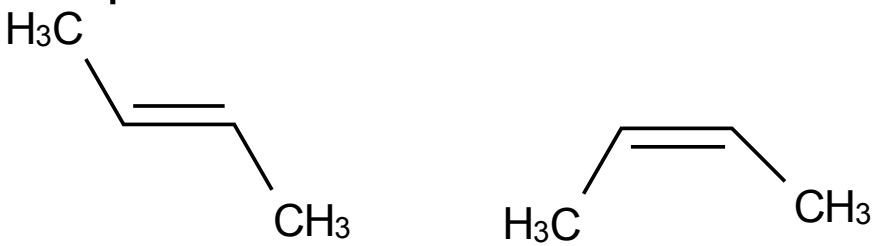
Химический поиск сложнее поиска подграфа

- Логические выражения
  SMARTS – SMiles ARbitrary Target Specification
- Ароматичность





- Стереохимия



Заключение

- Алгоритм поиска изоморфного подграфа
- Метод реверсивного поиска
- Алгоритм построение "фингерпринтов"и выбора кандидатов для подструктурного поиска
- Организация индекса

Ссылки на литературу:

[Cordella 2004] *A (sub)graph isomorphism algorithm for matching large graphs*; Cordella, L.P., Foggia, P., Sansone, C.,Vento, M., 2004

[Avis 1993] *Reverse Search for Enumeration*; David Avis, Komei Fukuda, 1993

[He 2006] *Closure-Tree: An Index Structure for Graph Queries*; Huahai He, Ambuj K. Singh, 2006

Дополнительная информация

Bingo – open-source chemistry search engine
(MS SQL Server, Oracle, Postgres?)

Контакты и ссылки:

- `http://ggasoftware.com/`
- `http://ggasoftware.com/opensource/bingo`
- rybalkin@ggasoftware.com
  michael.rybalkin@gmail.com