

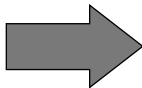
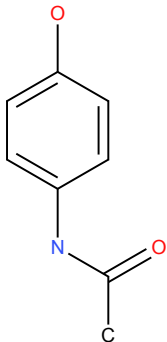
Распознавание изображений химических структур

Смолов Виктор, Зенцев Фёдор

2010

Задача

Восстановить машинное представление молекулы, утерянное при отрисовке



```
11 11 0 0 0 0
0.0000 3.3000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 2.4750 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.7145 2.0625 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 2.0625 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 1.2375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.4289 0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.4289 -0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 -0.4125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 -1.2375 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 0.8250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0 0 0 0
2 3 2 0 0 0 0
2 4 1 0 0 0 0
4 5 1 0 0 0 0
5 6 2 0 0 0 0
6 7 1 0 0 0 0
7 8 2 0 0 0 0
8 9 1 0 0 0 0
8 10 1 0 0 0 0
10 11 2 0 0 0 0
5 11 1 0 0 0 0
M END
```

Введение

Возникновение задачи: работа с публикациями

File View Molecule

Document Log Molecule

7 Synthesis of Porphyrins Fused with Aromatic Rings

Noboru Ono, Hiroko Yamada and Tetsuo Okujima

Department of Chemistry and Biology,
Graduate School of Science and Engineering,
Ehime University, 2-5 Bunkyo-cho, Matsuyama 790-8577, Japan

I. Introduction	1
II. Synthesis from Pyrroles	3
A. Template Synthesis	3
B. Isoindole Equivalents: Oxidative Method	8
C. Isoindole Equivalents: Retro-Diels-Alder Method	19
D. Use of Stable Isoindole Derivatives	44
III. Synthesis from Porphyrins	55
A. Bromoporphyrins: Metal-Catalyzed Reactions	58
B. Oxidative Coupling Reactions	65
C. From Nitroporphyrins	70
D. From Formylporphyrins	78
E. Diels-Alder Reactions	81
F. Condensation Reaction: Synthesis of Quinonoline-Fused Porphyrins	86
IV. Conclusion	95
V. Acknowledgment	96
VI. References	96

I. Introduction

The extension of the porphyrin π -system has been achieved by peripheral changes with different functional groups at the α - and β -positions. Fusion of aromatic rings at these positions is an effective method to control HOMO and LUMO energy levels of porphyrins. Such π -extended porphyrins have small HOMO-LUMO

1 / 102 75%

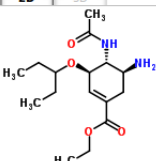
Введение

Возникновение задачи: пополнение баз знаний

Соединение не так интересно само по себе. Интересен контекст.

INHERENT PROPERTIES, IDENTIFIERS AND REFERENCES

2D 3D



ChemSpider ID: [58540](#) **Quick Links:** [Permalink](#) [Similar](#) !

Empirical Formula: [C₁₆H₂₈N₂O₄](#)

Molecular Weight: 312.4045

Nominal Mass: 312 Da

Average Mass: 312.4045 Da

Monoisotopic Mass: 312.204907 Da

Systematic Name: ethyl (3R,4R,5S)-5-ethoxy-1-methyl-4-oxocyclohex-1-en-3-ylideneacetate

Std. InChIKey: [VSZGPKBBMSAYNT-RRFJBJMHSA-N](#)

WIKIPEDIA ARTICLE(s)

ASSOCIATED DATA SOURCES AND COMMENTS

Chemical Vendors Biological Data
Phys. Properties Tox/Envir. Data Personal Data Web Article Data Aggregators

Каковы свойства?

Известны ли прекурсоры?

Есть ли патент?

В каких статьях упоминается?

Этапы:

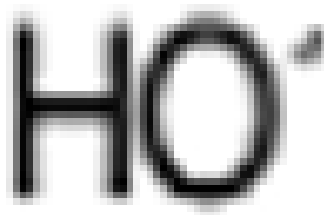
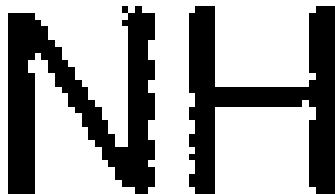
- 1 Обработка изображения
- 2 Разбор изображения
- 3 Извлечение связанных компонент
- 4 Отделение символьных компонент от графических
- 5 Извлечение графа
 - Кусочно-линейные элементы
 - Стереосвязи
 - Кольца
- 6 Распознавание символов
- 7 Объединение результатов
- 8 Исправление химических ошибок

Этапы распознавания

Обработка изображения

Призвана облегчить последующие этапы

- Фильтрация
- Бинаризация: перевод полутонового изображения в двухцветное

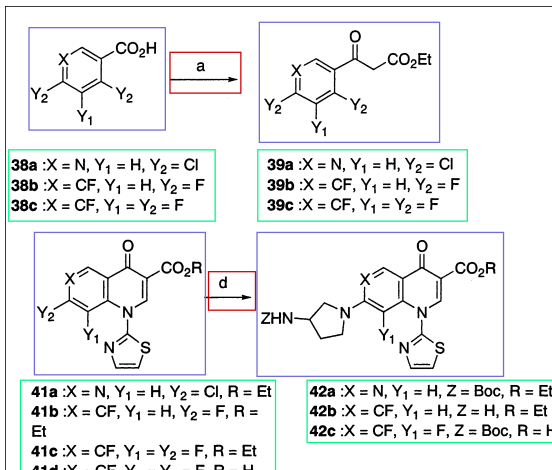


Этапы распознавания

Разбор изображения

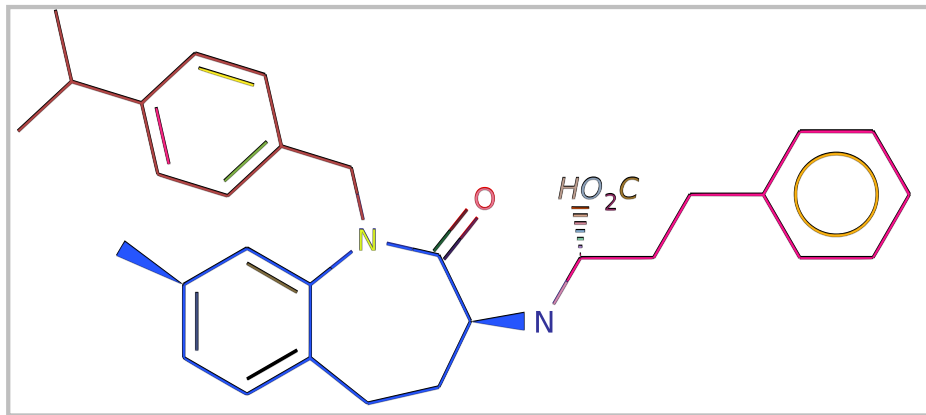
Изображение может быть сложным

- Таблицы заместителей
- Стрелки реакций
- Несколько молекул



Этапы распознавания

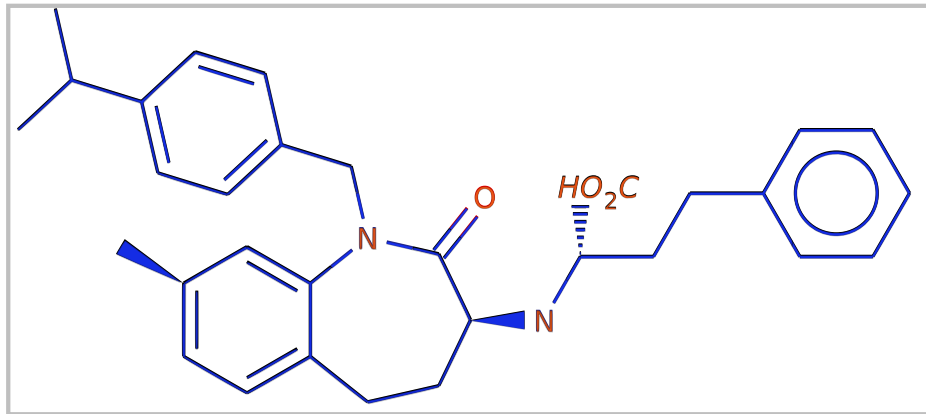
Сегментация. Разделение на компоненты связности



Этапы распознавания

Символьная и графическая информация

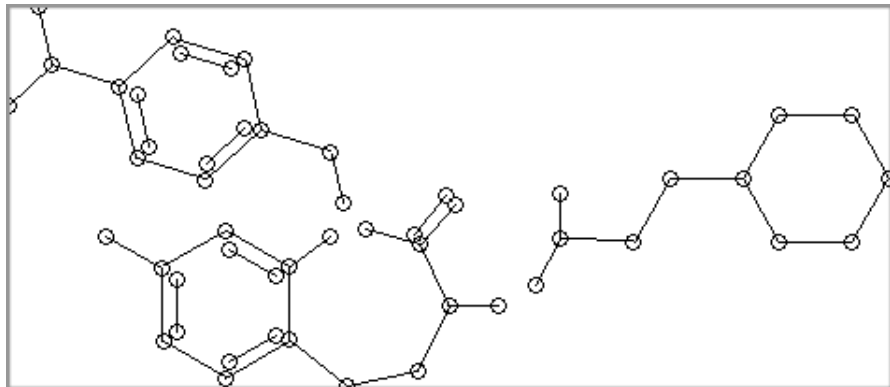
Разделение происходит на основе оценки высоты заглавного символа и различных эвристик



Этапы распознавания

Извлечение графа: кусочно-линейные элементы

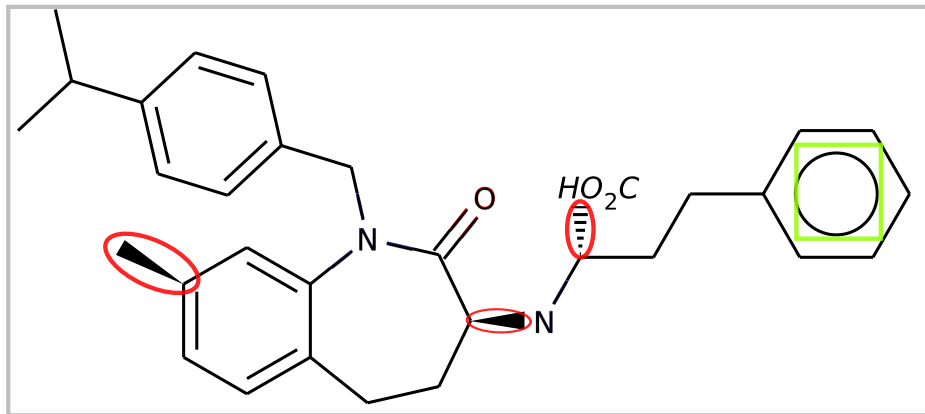
Картинка подвергается обработке фильтром утоньшения, а затем векторизации



Этапы распознавания

Извлечение графа: кольца и стереосвязи

- Распознавание стерео-вверх связей происходит после векторизации с помощью анализа толщины связи
- Кольца распознаются с помощью анализа дескрипторов контура сегмента



Этапы распознавания

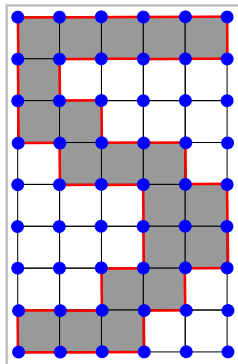
Распознавание символов

- 1 Получение контура обходом в глубину в специальном графе
- 2 Подсчет для контура дескрипторов Фурье

$$a_n = -\frac{1}{n\pi} \sum_{k=1}^m \Delta\phi_k \sin\left(\frac{2\pi n l_k}{L}\right)$$

$$b_n = \frac{1}{n\pi} \sum_{k=1}^m \Delta\phi_k \cos\left(\frac{2\pi n l_k}{L}\right)$$

- 3 Сравнение дескрипторов с эталонными



Этапы распознавания

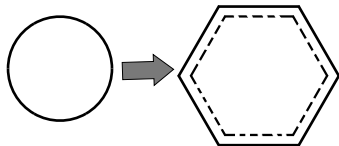
Объединение результатов и дополнение химической информацией

Построение молекулярного графа:

- Поиск и объединение кратных связей
- Объединение близких вершин
- Закрепление меток атомов

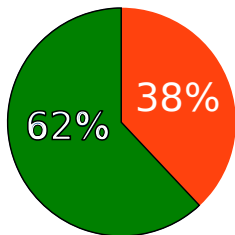
Дополнения:

- Ароматизация связей вокруг кольца
- Исправление неправильных направлений стереосвязей



Уже существуют несколько наборов изображений и соответствующих файлов с машинным представлением для испытаний

*USPTO*¹

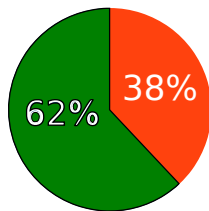


5700 изображений

~105 минут

~1.1 секунды на изображение

Chem-Infty



450 изображений

~10 минут

~1.3 секунды на изображение

¹United States Patent and Trademark Office

Известные

- 1 Фильтрация. Свертка изображения с различными ядрами.
- 2 Утоньшение
Joseph M. Cychosz, *Efficient Binary Image Thinning using Neighborhood Maps*, Graphics Gems IV, Academic Press, 1994
- 3 Распознавание символов
Charles T. Zahn & Ralph Z. Roskies, *Fourier Descriptors for Plane Closed Curves*, IEEE Transactions on computers, Vol. c-21, No. 3, march 1972

Оригинальные

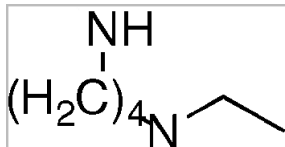
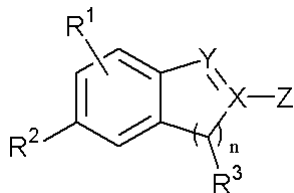
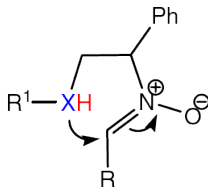
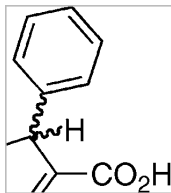
- 1 Разделение символьной и графической информации
- 2 Извлечение кусочно-линейных элементов

Планы на будущее

- Более сложная обработка изображения
- Учет дополнительной информации на картинке (таблицы заместителей, реакции)
- Раскрытие групп
- Интеллектуальный разбор изображения
- Контекстный анализ на основе известных из химии фактов
- Комплексный анализ документа
- Адаптация к работе с рукописными молекулам
- Улучшение качества распознавания

Заклучение

- Задача актуальна и требует решения
- Существует большое количество сложных подзадач
 - Слипшиеся связи и символы
 - Связи, изображаемые кривыми
 - Обобщенные молекулы



Спасибо за внимание!

<http://scitouch.net/imago>