

Хемоинформатика

Доклад на семинаре по
специальности

Студент гр.4057/2 Зенцев Фёдор
22 ноября

Введение

- Определение
- Хемоинформатика и другие науки

1 Фундаментальные вопросы

2 История

3 Представление химической информации

- Внутреннее
- Внешнее

4 Программные решения

5 Актуальные проблемы

Заключение

Использованные источники

Введение

Определение(1)

- Химическая информатика
- Хемоинформатика
- Chemical informatics
- Chem-, chemo- informatics

Определение (Ф.К.Браун, 1998)

Хемоинформатика означает совместное использование информационных ресурсов для преобразования данных в информацию и информации в знания для быстреего принятия наилучших решений при поиске соединений в разработке лекарств и их оптимизации

Введение

Определение(2)

Определение (Г.Пэриз, «Новартис»)

Хемоинформатика это научная дисциплина, охватывающая дизайн, создание, организацию, управление, поиск, анализ, распространение, визуализацию и использование химической информации

Определение (И.Гастайгер)

Хемоинформатика это применение методов информатики для решения химических проблем

Определение (Р.Грин)

Хемоинформатика - новое название старых проблем

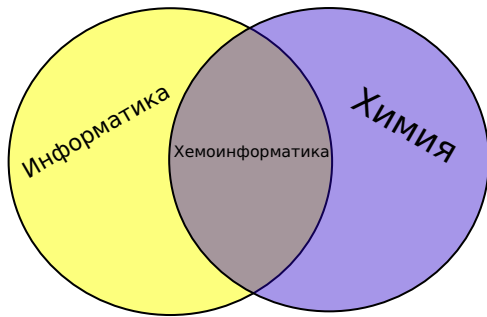
Введение

Хемоинформатика и другие науки(1)



Введение

Хемоинформатика и другие науки(2)



Основные используемые инструменты информатики и математики:

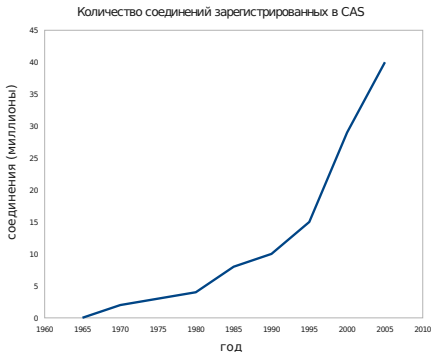
- Базы данных
- Интеллектуальный анализ данных
- Математическая статистика
- Машинное обучение
- Теория графов
- Компьютерная графика

- 1 Какое **химическое соединение** обладает интересующим свойством?
Свойство → Структура
- 2 Как получить такое **химическое соединение**?
Исходные вещества → Соединение
- 3 Какое **химическое соединение** получится в результате реакции?
Реакция → Продукт



Причины появления хемоинформатики

- Огромное количество информации: миллионы соединений, миллионы публикаций
- Сложные химико-биологические связи
 - Потребность в математических расчетах
 - Потребность в визуализации
- Низкая результативность экспериментов
- Дороговизна натуральных экспериментов
- Вечный поиск новых лекарств

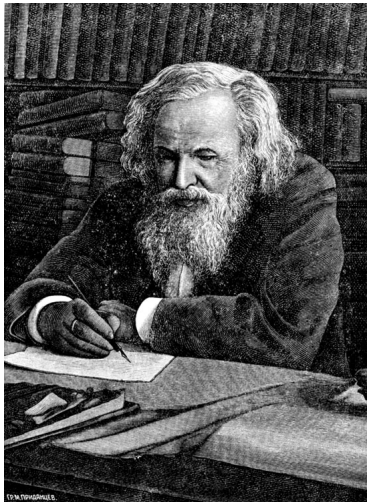


¹CAS: Chemical Abstracts Service, химическая реферативная служба

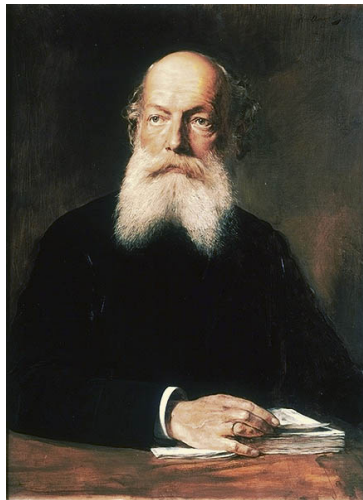
История

Первые «хемоинформатики»

Дмитрий Иванович Менделеев



Фридрих Август Кекуле

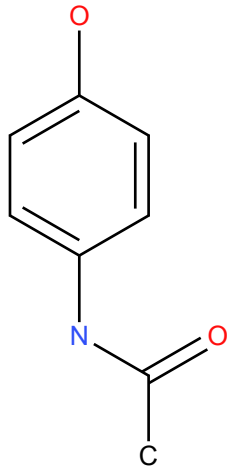


- Базы данных
 - 1957 год. Вашингтон, Национальное Бюро Стандартов
 - 1960 год. Начало финансирования Химической Реферативной Службы для составления баз данных и разработки методов поиска
- Визуализация
 - Конец шестидесятых. Принстон и Сан-Франциско
- Разработка синтеза
 - 1969 год. Гарвард
- Структурное разъяснение (*elucidation*) и генерация химических соединений
 - 1964 год. Стэнфорд, проект DENDRAL
 - 1970 год. Университет Аризоны

Представление химической информации

Внутреннее: модель молекулярного графа(1)

Химия	Теория графов
Молекула	Связный компонент
Атом	Вершина
Связь	Ребро
Название атома	Метка вершины
Порядок связи	Метка ребра
Валентность атома	Степень вершины



Изоморфизм структурной формулы молекулы и молекулярного графа

Представление химической информации

Внутреннее: модель молекулярного графа(2)

Химия	Теория графов
Свойство молекулы, дескриптор молекулы	Инвариант графа
Одинаковые молекулы	Изоморфные графы
Наличие определенного фрагмента в молекуле	Изоморфизм графа подграфу
Поиск подструктуры	Поиск изоморфного подграфа
Пересечение молекул	Максимальный общий подграф
Порядок связи	Метка ребра
Топологическая группа симметрий	Группа автоморфизмов графа

Представление химической информации

Внешнее: химические форматы файлов(1)

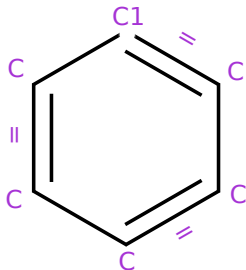
■ Строковое представление, SMILES¹

- Метан $CH_4 \xrightarrow{SMILES} C$
- Этанол $C_2H_6O \xrightarrow{SMILES} CCO$
- Бензол $C_6H_6 \xrightarrow{SMILES} C1=CC=CC=C1$

■ Строковое представление, InChi²

- Этанол
 $C_2H_6O \xrightarrow{InChi} InChi =$
 $1/C2H6O/c1-2-3/h3H,2H2,1H3$

■ Представление с помощью языка разметки, CML³



²Simplified Molecular Input Line Entry Specification

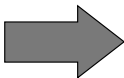
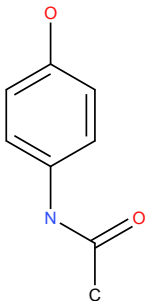
³IUPAC International Chemical Identifier

⁴Chemical Markup Language

Представление химической информации

Внешнее: химические форматы файлов(2)

- Формат XYZ = метки атомов + их координаты. Отражает геометрию молекулы.
- Табличное представление, MDL⁴Molfile
- Табличное представление, SDF⁵



```
par acet ombl um
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
```

⁵Molecular Design Limited, Inc.

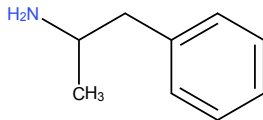
⁶Structure-Data File

Программные решения

Базы данных: химические соединения

Химический картридж

- Поисковый движок для базы данных химических соединений
- Программа, расширяющая функциональность СУБД и позволяющая работать с ней в терминах предметной области



Возникают сложные задачи теории графов

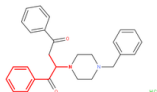
Response time: 12.42 sec

Results: 8 of 3584

1 2 3 4 5 6 7 8

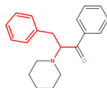
ID: 270

SMILES: N(CCC1=CC=CC=C1)SCCN(C(C)C(=O)C2C=CC=CC=C2)C(=O)C3=CC=CC=C3C2(C)C1



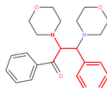
ID: 279

SMILES: N(C(C)C(=O)C1=CC=CC=C1)CC(C=CC=CC=C1)SCCCC1



ID: 312

SMILES: O1CCN(C(C)C2C=CC=CC=C2)C(=O)C3C(NC(C)C(C)C2)C2=CC=CC=C2C1

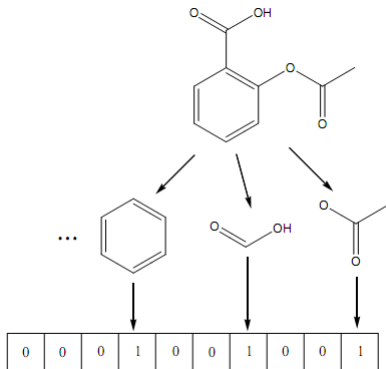


Программные решения

Поиск количественных соотношений структура-свойство (QSAR⁶)

Отображение химического пространства на пространство дескрипторов

- Фрагментные бинарные и целочисленные дескрипторы или структурные ключи
- Физико-химические дескрипторы
- Квантово-химические дескрипторы
- Топологические индексы
- Дескрипторы молекулярных полей



⁷QSAR: Quantative structure-activity relationship

Программные решения

Молекулярные дескрипторы(1)

- Дескрипторный подход - другое представление химической информации, неграфовое.
- При прогнозировании свойств различают две задачи:
 - Классификационная - качественный уровень
 - Регрессионная - числовые значения

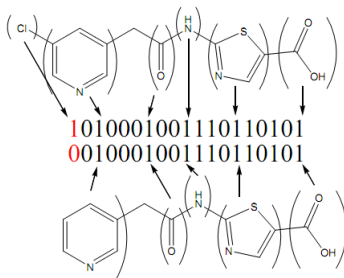
Применяются методы математической статистики и машинного обучения

Программные решения

Молекулярные дескрипторы(2)

Дескрипторы применяются не только для прогнозирования свойств соединений

- Ускорение подструктурного поиска
- Задача молекулярного подобия
- Нахождение одинаковых соединений



$$T = \frac{N_{A \& B}}{N_A + N_B - N_{A \& B}} = 0.9$$

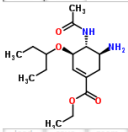
Программные решения

Базы данных: публикации

Найденные в базе данных молекулы не очень интересны сами 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)-4-(acetylamino)-5-amino-3-(pentan-3-yloxy)cyclohex-1-ene

SMILES: O=C(OCC)/C1=C/C([C@@H](OC(CC)CC)[C@H](NC(=O)C)[C@@H](N)C1 [Copy](#)

InChI: InChI=1/C16H28N2O4/c1-5-12(6-2)22-14-9-11(16(20)21-7-3)8-13(17)15(14)115H,5-8,17H2,1-4H3,(H,18,19)/t13-,14+,15+/m0/s1 [Copy](#)

InChIKey: VSZGPKBBMSAYNT-RRFJBMHBB

Std. InChI: InChI=1S/C16H28N2O4/c1-5-12(6-2)22-14-9-11(16(20)21-7-3)8-13(17)15(14):10(4)19/h9,12-15H,5-8,17H2,1-4H3,(H,18,19)/t13-,14+,15+/m0/s1 [Copy](#)

Std. InChIKey: VSZGPKBBMSAYNT-RRFJBMHSA-N

WIKIPEDIA ARTICLE(s)

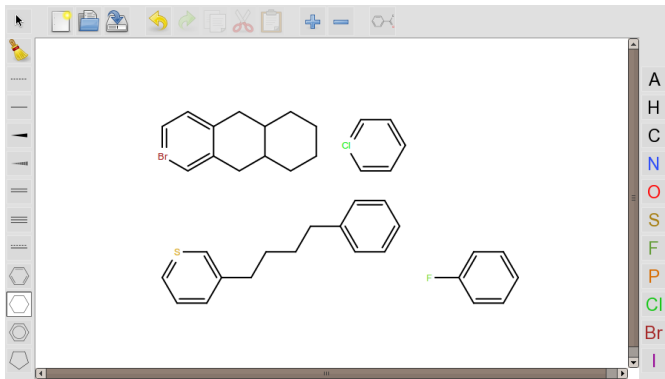
ASSOCIATED DATA SOURCES AND COMMERCIAL SUPPLIERS

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

Возникают задачи интеллектуального анализа данных

Программные решения

Редакторы молекул

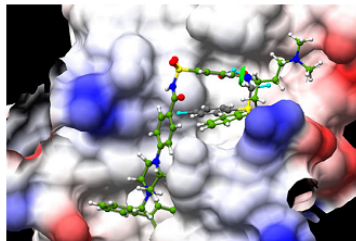


- Возникают задачи укладки графа на плоскости, создание общего стандарта отрисовки, эстетики рисунка

Программные решения

Докинг или молекулярная стыковка

- Метод, позволяющий предсказать наиболее выгодную для создания устойчивого комплекса ориентацию и положение одной молекулы относительно другой



Возникают задачи оптимизации

- *in silico* contra *in vivo* et *in vitro*
- Некачественное программное обеспечение для исследователей
 - Пользовательские интерфейсы
 - Производительность
- Разобщенность обладателей баз данных, отсутствие стандартов



- Хемоинформатика - сравнительно молодая область
- Роль информатики в химии и фармакологии высока уже сейчас
 - Дороговизна натуральных экспериментов
 - Очень долгий срок разработки лекарства
- Химия нуждается в информатике и математике
 - Создаются университетские специальности по хемо- и биоинформатике
 - Проводятся конференции

- 1 J.Gasteiger, T.Angel *Chemoinformatics: A Textbook*, 2003
- 2 F.K.Brown *Chemoinformatics: What is it and How does it Impact Drug Discovery*, 1998
- 3 B.A.Bunin, B.Siesel, G.A.Morales, J.Bajorath *Chemoinformatics: Theory, Practice, & Products*, 2007
- 4 A.R.Leach, V.J.Gillet *An introduction to chemoinformatics*, 2007

Использованные источники

Журналы, презентации, интернет

- 1 Дмитрий Павлов, *Навигация в мире органических соединений*, Компьютерные инструменты в образовании, 2010, №3
- 2 Сергей Кокорин, *Заметки о Cheminfo'S. Strasbourg Summer School on Chemoinformatics*.
- 3 Материалы AACIMP-2008, курс “Хемоинформатика”
<http://summerschool.ssa.org.ua/>
- 4 Noel O'Boyle, <http://baoilleach.blogspot.com/>
- 5 Википедия (русская, английская)

Спасибо за внимание!