

Лекция 2. Поиск новых биоактивных молекул и хемоинформатика

Курс: Молекулярное моделирование в применении к биомолекулам

Головин А.В.¹

¹МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики

Москва, 2020

Содержание

Активные молекулы

Фарминдустрия

HTS

Хемоинформатика

QSAR



Активные молекулы

- В основном биологически активные молекулы взаимодействуют нековалентно с биополимерами
- Агонисты связываются как нативные лиганды и дают тот же эффект
- Антагонисты конкурируют или препятствуют связыванию нативного лиганда
- Обратные агонисты связываются и оказывают эффект, обратный эффекту нативного лиганда
- Хорошие молекулы показывают высокую комплементарность поверхности биополимера



Свойства лекарства

- Лекарством обычно являются не только те молекулы, которые хорошо связываются с биополимером.
- Лекарство должно иметь приемлемую растворимость
- Часто бывает, что лекарству надо проникнуть сквозь мембрану.
- Хорошо когда лекарство в итоге метаболизируется, а не накапливается в тканях.



Как искать активные молекулы?

- Можно пытаться искать вещества в биоматериалах.
- Можно проводить роботизированное сканирование библиотеки соединений на активность в разных тестах.
- Недостаток сканирования: не все тесты можно адаптировать под робота.
- Возможен высокий уровень шума из-за не специфических взаимодействий
- Можно применить фильтрацию по подобию соединений, для этого нужны ИТ.



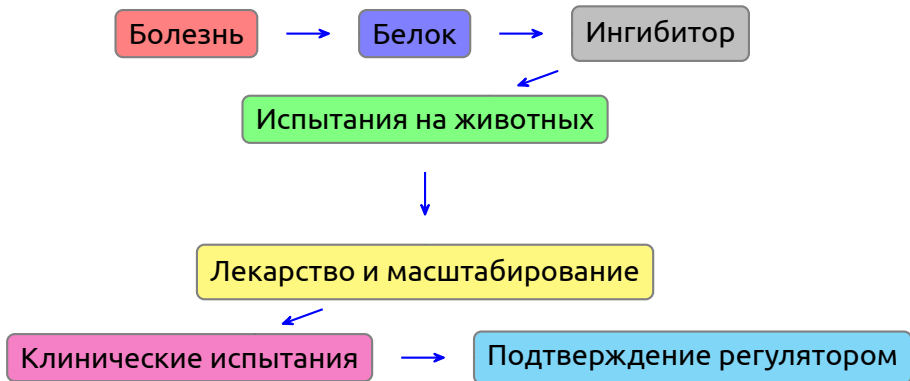
Особенности деятельности фарм-производителей

Дженерик - лекарство без патентной защиты (срок вышел)

- Рынок высоко конкурентен.
- Разработка нового лекарства занимает от 10 до 20 лет.
- Новые лекарства приносят основную прибыль
- 4 основные фазы: открытие, разработка, испытания, продажи



R&D



Новые технологии

- Чипы: экспрессия генов.
- Структуры: роботизированный поиск комплексов с кристаллом белка.
- Высоко-производительный поиск ингибиторов.
- Виртуальный поиск.
- Комбинаторная химия.

Все это в основном относится к стадии поиска ингибитора

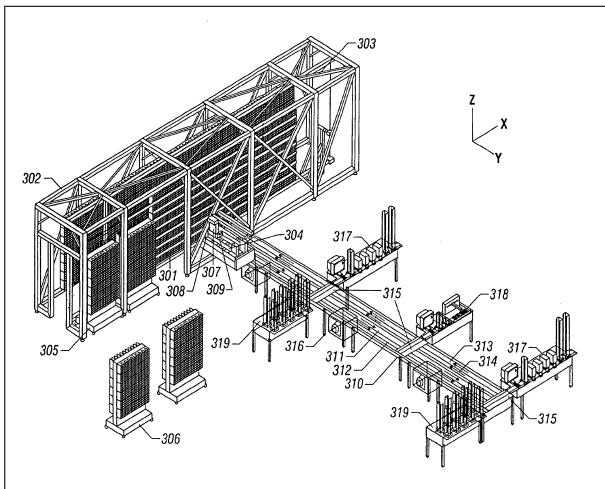


Как хемоинформатика может помочь?

- Разработка методов и управление информацией о лигандах.
- Оценка данных *in silico* для минимизации рисков.
 - Разработка библиотеки.
 - Виртуальный поиск.
 - Оценка стоимости и выгоды.
- Организация доступа к информации.
- Интеграция процессов.



Пример: HTS, Высоко-производительный поиск ингибиторов



до 100000 соединений в день

HTS и поток данных

- Исполнить HTS.
- Решить какие соединения активны а какие нет.
- Кластеризация активных соединений в классы.
- Визуализация.
- Идентификация "основы" для каждого класса.
- Поиск причин, элементов структуры, которые приводят к "не активности".
- Использование структурной информации для объяснения активности.



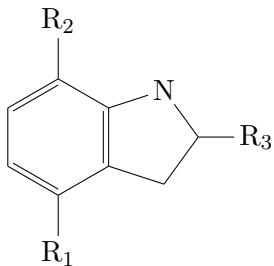
Пример, комбинаторная химия

- Исследователи используют "строительные блоки" для быстрого создания большого количества разных соединений.
- Обычно используется некоторая "основа" и "строительные блоки" присоединяются к разным местам основы.



Комбинаторная химия

"Основа"



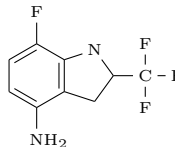
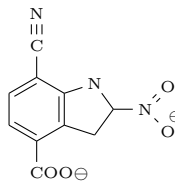
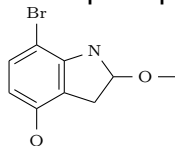
"Блоки"

$R_1 = \text{OH}, \text{OCH}_3, \text{NH}_2, \text{Cl}, \text{COOH}$

$R_2 = \text{Phe}, \text{OH}, \text{NH}_2, \text{Br}, \text{F}, \text{CN}$

$R_3 = \text{CF}_3, \text{NO}_2, \text{OCH}_3, \text{OH}, \text{PheO}$

Примеры



Хемоинформатика и библиотеки

- Какие блоки выбрать?
- Какие библиотеки строить?
 - Дополнение известных наборов
 - Модификация под конкретный белок
 - Полное "насыщение" библиотеки
- Компьютерное профилирование библиотеки
 - Виртуальными библиотеками удобно манипулировать на компьютере



Компьютерное представление молекул

- Хранение в компьютере молекулы как изображения имеет малую ценность
- Большинство современных баз данных представляет молекулу как граф, с узлами и рёбрами
- Графы представляются как таблицы связей.

Marvin 04200617372D

```

4 3 0 0 0 0          999 V2000
  0.0000  0.0000  0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0.7145  -0.4125  0.0000 D  0 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
  0.0000  0.8250  0.0000 D  0 0 0 0 0 0 0 0 0 0 0 0 0
1  4  2  0  0  0  0
2  1  1  0  0  0  0
3  1  1  0  0  0  0
M  END

```

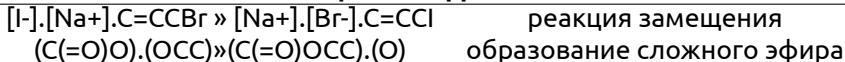


Линейное представление молекул, SMILES

Молекула представляется в виде диаграммы и каждый атом проходится только один раз

<chem>CC</chem>	ethane	<chem>[OH3+]</chem>	hydronium ion
<chem>O=C=O</chem>	carbon dioxide	<chem>[[2H]]O[[2H]]</chem>	deuterium oxide
<chem>C#N</chem>	hydrogen cyanide	<chem>[[235U]]</chem>	uranium-235
<chem>CCN(CC)CC</chem>	triethylamine	<chem>F/C=C/F</chem>	E-difluoroethene
<chem>CC(=O)O</chem>	acetic acid	<chem>F/C=C/F</chem>	Z-difluoroethene
<chem>C1CCCCC1</chem>	cyclohexane	<chem>N[[C@@H]](C)C(=O)O</chem>	L-alanine
<chem>c1ccccc1</chem>	benzene	<chem>N[[C@H]](C)C(=O)O</chem>	D-alanine

Реакции в виде SMILES



Стандартизация SMILES

- Очевидно, что одну молекулу можно описать разными способами.
- Морган в 1965 году предложил рассматривать каждый атом по свойству его окружения.
- Стандартные SMILES называют Unique.

Input SMILES	Unique SMILES
<chem>OCC</chem>	<chem>CCO</chem>
<chem>[CH3][CH2][OH]</chem>	<chem>CCO</chem>
<chem>C-C-O</chem>	<chem>CCO</chem>
<chem>C(O)C</chem>	<chem>CCO</chem>
<chem>OC(=O)C(Br)(Cl)N</chem>	<chem>NC(Cl)(Br)C(=O)O</chem>
<chem>ClC(Br)(N)C(=O)O</chem>	<chem>NC(Cl)(Br)C(=O)O</chem>
<chem>O=C(O)C(N)(Br)Cl</chem>	<chem>NC(Cl)(Br)C(=O)O</chem>



Описание SMILES: атомы

- Одно буквенные атомы, а именно : B, C, N, O, P, S, F, Cl, Br, I записываются как есть, как один символ.
- Все остальные атомы записываются в квадратных скобках [Pt]
- Так как атомы водорода обычно не указываются, то “валентность” атомов определятся как наименьшая из ближайших т.е. B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6).
- “Валентности”, отличные от “нормальных”, указывают в скобках [S], [H+], [Fe+2], [OH-], [Fe++], [OH3+], [NH4+]



Описание SMILES: связи

CC	этан
C=C	этилен
O=C=O	CO ₂
C#N	HCN
CCO	этанол
[H][H]	водород

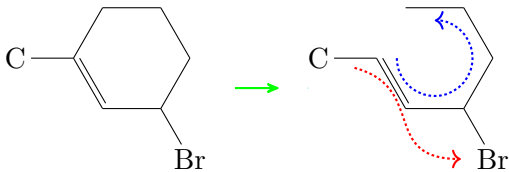
Ветвление цепи отображается в скобках ()

Пример: CCC(CC)COO



Описание SMILES: циклы

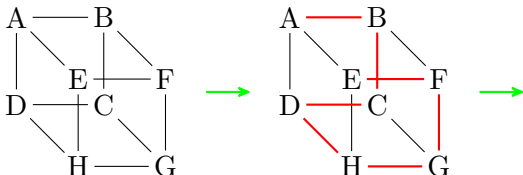
- C1CCCCC1 циклогексан



a) CC1=CC(Br)CCC1

b) CC1=CC(CCC1)Br

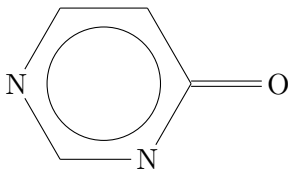
- Или более сложный пример:



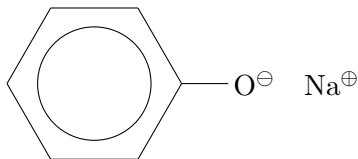
A14B2C3D1H5G3F2E45

Описание SMILES: ароматика

- SMILES для определения ароматичности использует расширенный алгоритм Хюккеля.
- c1ccccc1 eq C1=CC=CC=C1 тут все атомы находятся в sp^2 -гибридизации
- c1cccc1 eq C1=CC=CC1 , последний атом в гибридации sp^3 .
- Ароматичными могут быть атомы: C, N, O, P, S, As, Se, и *.
- Пример: c1nc[nH]c(=O)1



Структуры где есть нековалентные связи



В SMILES нотации это:

[Na+].[O-]c1ccccc1

или

c1cc([O-].[Na+])ccc1



Изомеры

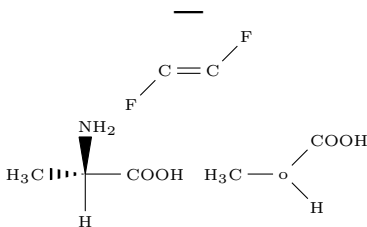
Изотопы

[12C],[13C]

Цис-Транс

F/C=F/C или F\C=F\C

Хиральность



N[C@](C)(C(=O)O)H
N[C@@](C)(H)C(=O)O



SMARTS: паттерны для SMILES

В принципе, SMARTS это SMILES + операторы логики и варианты в позициях.

Пример для атомов:

C	алифатический углерод
c	ароматический углерод
a	любой ароматический атом
[#6]	любой атом углерода
[++]	атом с зарядом +2
[R]	атом в кольце
[D3]	атом с тремя связями (не с атомами водорода)
[X3]	атом с тремя связями, включая атомы водорода
[v3]	атом с валентностью 3.



SMARTS: логические операторы и примеры

Логика:

!e1	not e1
e1& e2	a1 and e2
e1,e2	e1 or e2
e1;e2	a1 and e2

Пример:

[!C;R]	не алифатический C в кольце
[n;H1], [n&H1], [nH1]	H в пирроле
[c,n&H1]	C или H в пирроле
[X3&H0]	Атом с тремя связями не с H
[c,n;H1]	N или C в связи с одним H1

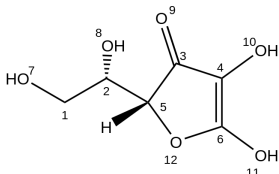


Линейное представление молекул, InChI

InChI = IUPAC International Chemical Identifier

Структура молекул описывается слоями :

- Основной слой содержит описание брутто формулы, связанности (c) и связей с водородами (h)
C₂H₆O/c1-2-3/h3H,2H₂,1H₃
- Слой с описанием заряда (p) кратности связей
- Слой с описанием стереохимии и связей
C₆H₈O₆/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H₂/t2-,5+



Дискрипторы, правило Лепински

- Водородные связи
- Гибкость молекулы
- Гидрофобность

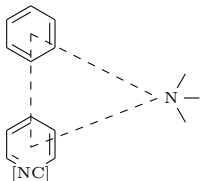
Правило пяти Лепински

- No more than 5 hydrogen bond donors
- No more than 10 hydrogen bond acceptors
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient $\log P$ not greater than 5



Поиск по 3D-базам данным

- Поиск в 2D-пространстве хорош для поиска подобных молекул, но биологически активные молекулы действуют благодаря специфической 3D-структуре.
- Взаимодействие с биополимером может происходить благодаря нужному расположению в пространстве некоторых групп. При этом различие в 2D-структуре может быть весьма существенным.
- Фармакофор — это набор свойств, которые являются общими для некоторой группы активных молекул.
- Пример: Антигистаминный 3D-фармакофор



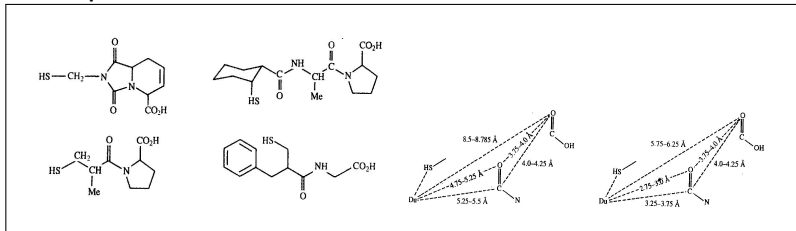
Проблемы с фармакофорами

- Если молекулы более или менее подвижны, то это накладывает дополнительные требования на учёт конформационных превращений.
- Для определения фармакофора надо определить, какой набор групп располагается в биополимере идентично.
- Надо быть уверенным, что выбранный набор молекул связывается с белком в одном и том же месте. Однозначное указание на это можно получить только экспериментально.



Систематический поиск

- Есть проблема:



- Выбирают точки, которые по мнению исследователей определяют активность. Делают конформационный поиск для всех молекул. Если находят пересечения по геометрии, то на основе этих точек и геометрии пересечения формулируют фармакофор.

Базы данных:

- PubChem
- Cambridge database
- Inorganic structural database

The screenshot displays the PubChem Structure Search interface. On the left, the search options are visible, including 'Name/Text', 'Identity/Similarity', 'Substructure/Superstructure', 'Molecular Formula', and '3D Conformer'. The 'Identity/Similarity' option is selected. Below this, there is a 'Draw a Structure' section with a text input field for 'Enter single structure identifier' and buttons for 'Edit', 'Search', and 'Preview'. There are also sections for 'Options' and 'Filters'.

On the right, the 'Ibuprofen - Compound Summary (CID 3672)' is displayed. It includes the molecular formula $C_{13}H_{18}O_2$ and a 2D chemical structure. Below the structure, there is a 'Table of Contents' and an 'Identification' section listing various synonyms for Ibuprofen, such as 'Ibuprofen', 'Brufen', 'Nurofen', and 'Dolipran'.


```
In [1]: import pubchempy as pcp
```

```
In [21]: name='Aspirin'  
list=pcp.get_compounds(name, 'name')  
asp=list[0]  
hb=asp.h_bond_acceptor_count  
form=asp.molecular_formula  
xlogp=asp.xlogp  
print name, ":", form, " hb acceptors: ", hb, " Hydrophobicity: ", asp.xlogp
```

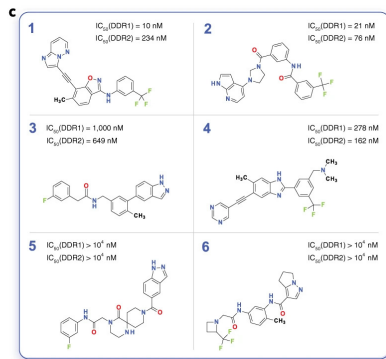
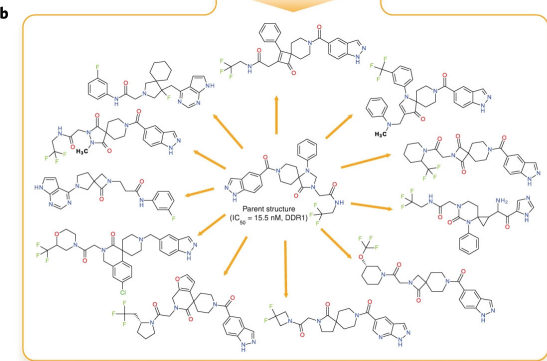
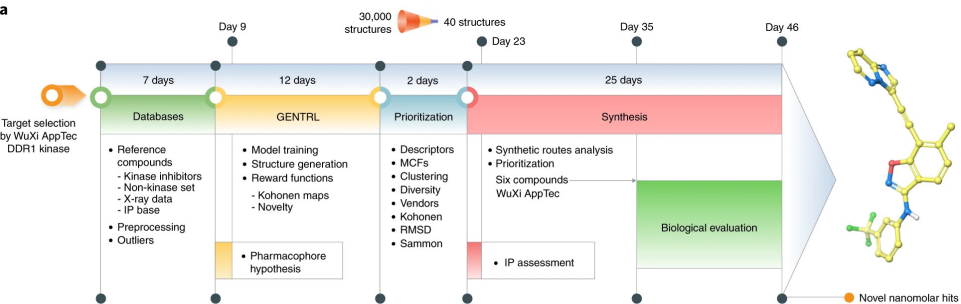
Aspirin: C9H8O4 hb acceptors: 4 Hydrophobicity: 1.2

```
In [30]: list=[]  
for i in range(1,10):  
    try:  
        a=pcp.get_compounds('C1=N-C=C-N1', 'smiles', searchtype='similarity', listkey_count=50, listkey_start=i)  
    except:  
        print "ended on: ", i*50  
        break  
    list.extend(a)  
  
print "Downloaded: ", 50*len(list), " compounds"
```

Downloaded: 22500 compounds

```
In [31]: list
```

```
Out [31]: [Compound(12749),  
Compound(82140),  
Compound(484),  
Compound(96125),  
Compound(283401),  
Compound(559542),  
Compound(2773261),  
Compound(2773328),  
Compound(135136),  
Compound(10080473),  
Compound(44345402),  
Compound(444230),  
Compound(444234),  
Compound(444235),  
Compound(444800).
```



Направления для работы

- Методы основанные на подобии рассматривают подобные вещества и белки, равномерное распределение отсутствует.
- Описание features сделать количественным.
- Методы основаны на datasets. Нужна адаптация под успешные предсказания.



Направления для работы

- Объединение баз данных. Комбинирование максимально доступного количества данных для пары белок-ингибитор.
- Правильно включение структурно-функциональных данных для лигандов и белков.

