Поиск новых биоактивных молекул и химоинформатика

Структурная Биоинформатика (МФК)

Головин А.В. [1]

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

Москва, 2013

Содержание

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?

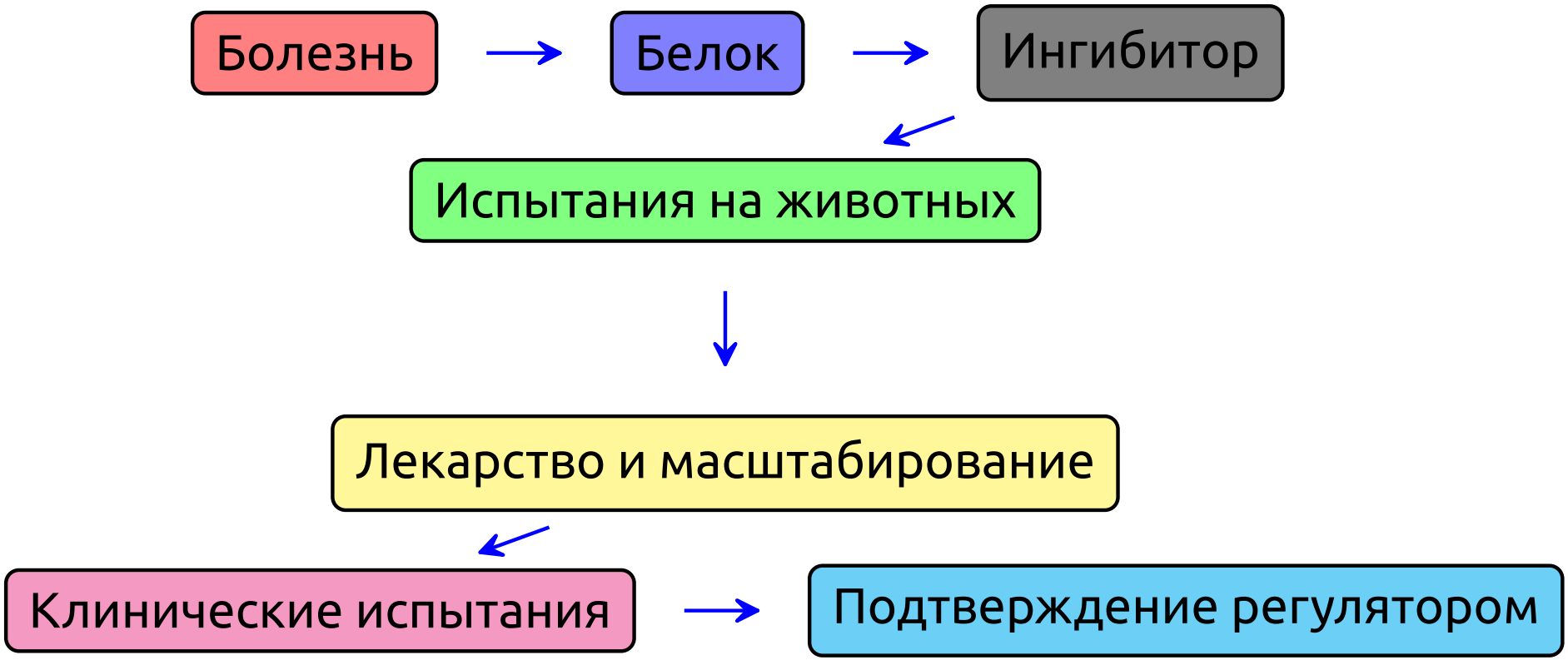
- Можно пытаться искать вещества в биоматериалах.
- Можно проводить роботизированное сканирование библиотеки соединений на активность в разных тестах.
- Недостаток сканирования: не все тесты можно адаптировать под робота.
- Возможен высокий уровень шума из-за неспецифических взаимодействий
- Можно применить 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


Болезнь
Белок
Ингибитор
Испытания на животных
Лекарство и масштабирование
Клинические испытания
Подтверждение регулятором

Новые технологии

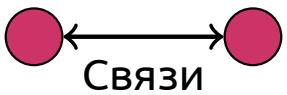
- Чипы: экспрессия генов.
- Структуры: роботизированный поиск кристаллов.
- Высоко-производительный поиск ингибиторов.
- Виртуальный поиск.
- Комбинаторная химия.

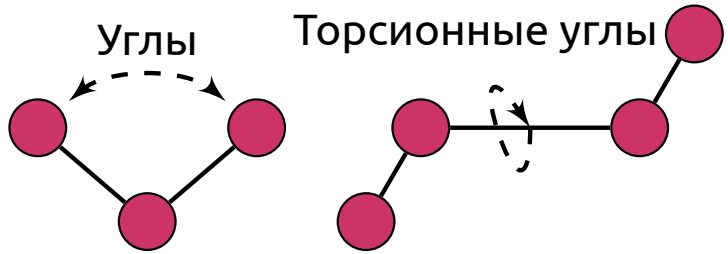
Все это в основном относится к стадии поиска ингибитора

Простое уравнение силового поля (СП)

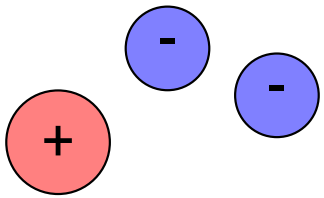
$$U = \sum_{bonds} \frac{k_i}{2}(l_i - l_0)^2 + \sum_{angles} \frac{k_i}{2}(\phi_i - \phi_0)^2 + \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma)) +$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)$$

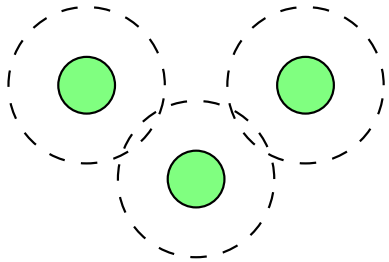




Ван-дер-Ваальсовы взаимодействия



Кулоновские взаимодействия



Простое уравнение силового поля (СП)

$$U = \sum_{bonds} \frac{k_i}{2}(l_i - l_0)^2 + \sum_{angles} \frac{k_i}{2}(\phi_i - \phi_0)^2 + \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma)) +$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)$$

Молеклярная динамика

Монте-Карло

$$v(t + \frac{\Delta t}{2}) = v(t - \frac{\Delta t}{2}) + \frac{F(t)}{m}\Delta t$$

$$acc\,(o \to \eta) = min\left(1, exp\left\{-\beta\left[U(r^N) - U(r^N)\right]\right\}\right)$$

Метод Монте-Карло

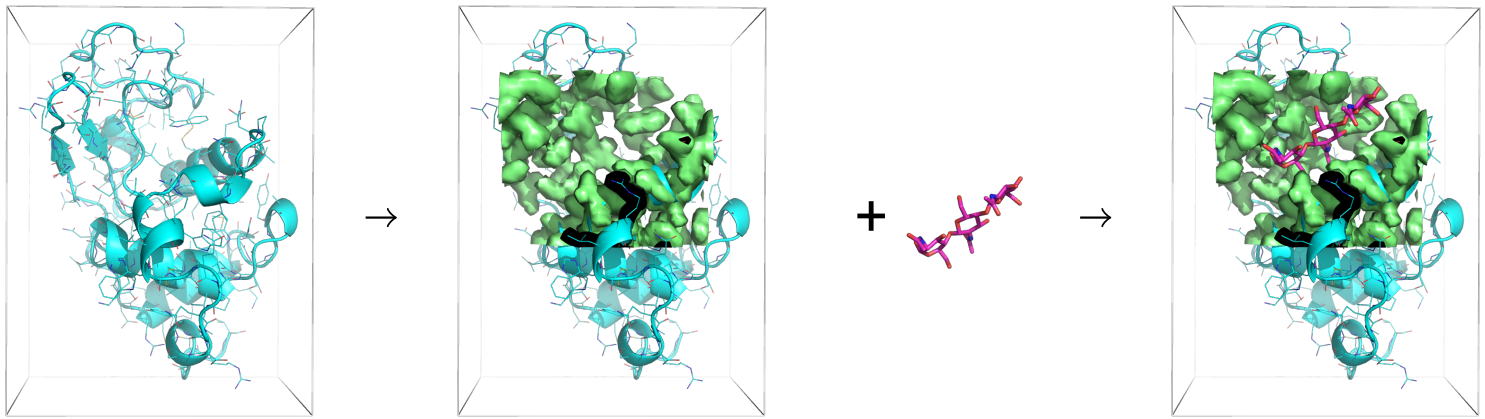
- Основная идея предполагает поиск конформаций с низкой энергией на основе случайного изменения координат.
- Скорости у атомов не рассчитываются
- Неудобно, что основной счёт приходит на ненужные состояния.
- Метрополис и соавторы предложили использовать цепи Маркова для генерации конформаций.

Пример:

1. генерируем: $x_{new} = x_{old} + (2\xi - 1)\delta r_{max}$
2. если новая энергия меньше старой то принимаем конформацию и используем дальше
3. если энергия выше то сравниваем фактор Больцмана со случайным числом 0:1 и если фактор Больцмана изменения энергии больше, то принимаем конформацию

Докинг белок-лиганд

Метод поиска способа связывания лиганда с белком



Монте-Карло моделирование в решётке, которая описывает окружение лиганда
В результате мы можем узнать положение лиганда в комплексе с белком и оценить константу связывания

Положение в сайте связывания

- Сайт связывания — место связывания лиганда
- Геометрия связывания — место связывания, ориентация и конформация лиганда

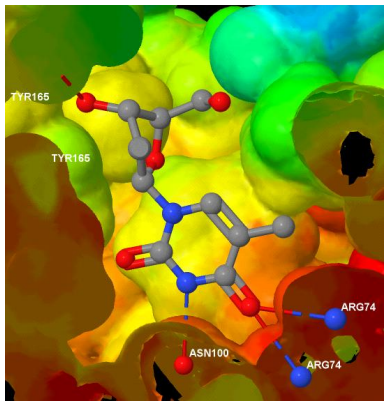
Использование докинга

Основные цели докинга:

- Виртуальный поиск лигандов
- Определение геометрии связывания лиганда

Если мы знаем, как связывается лиганд, то:

- Мы можем узнать, какие части важны для связывания
- Можно предложить изменения для улучшения константы связывания
- Можем избежать ошибок



Два основных компонента программ для докинга

- Алгоритм поиска
  - Установление места связывания
  - Установление геометрии связывания
- Алгоритм расчёта константы связывания областей с низкой энергией.

Реализация

Сегодня существует много программ для докинга

- AutoDock, DOCK, e-Hits, FlexX, FRED, Glide, GOLD, LigandFit, QXP, Surflex-Dock…и т.д.
- разные алгоритмы оценки аффинности и разные алгоритмы поиска
- Важно не путать лиганд-белок докинг и белок-белок докинг

Практические аспекты

- Часто PDB-структура содержит молекулы воды, почти всегда их надо убрать.
- Надо добавлять протоны к структуре; His?
- Часто в PDB неточно определена ориентация некоторых групп, что сказывается на паттерне водородных связей.
- Протонирование лиганда и его таутомерные формы.

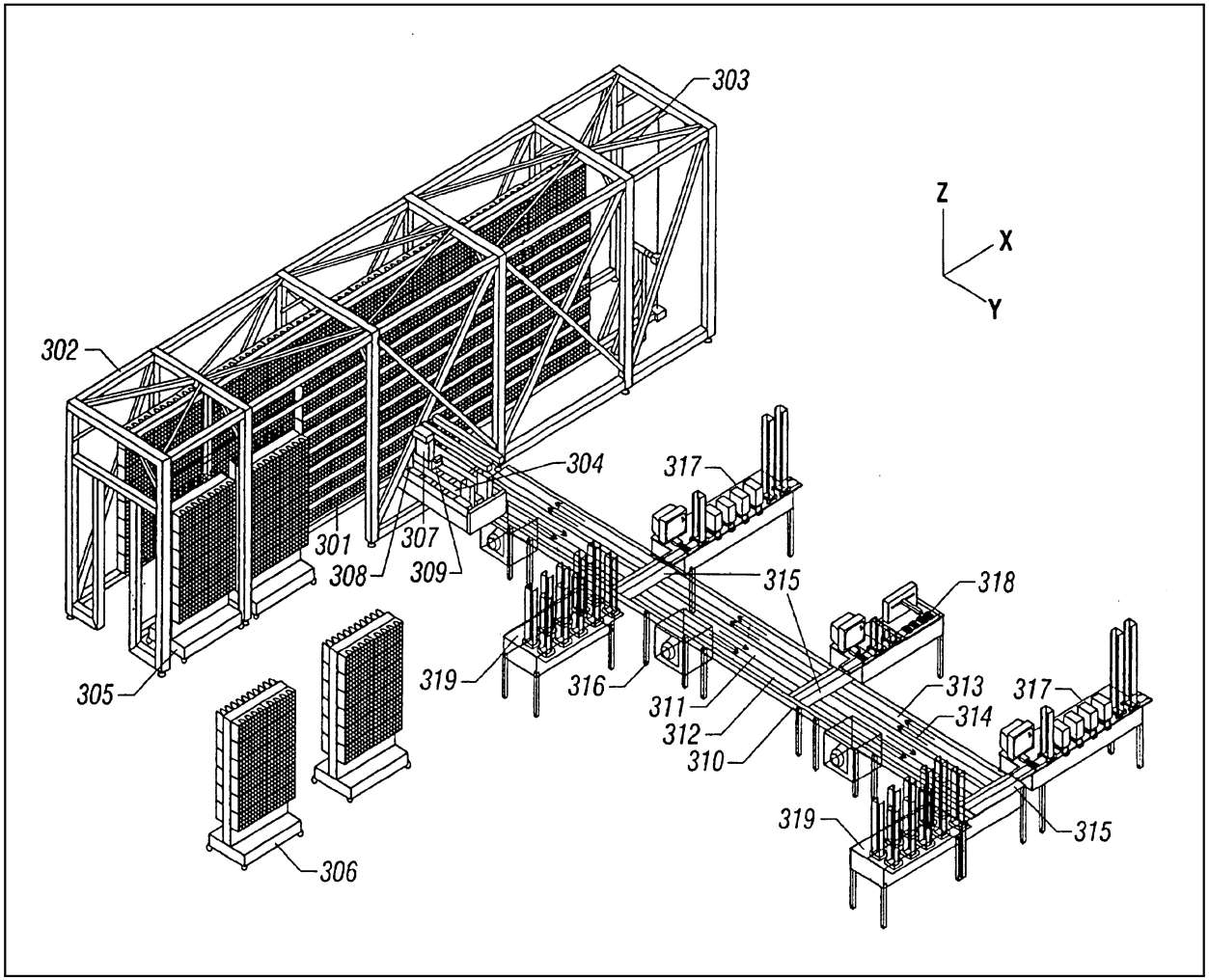
Rigid|Flexible докинг

- Rigid: лиганд не имеет внутренних степеней свободы, т.е. вращение вокруг связей запрещено.
- Flexible: предполагает учёт вращения вокруг связей лиганда.
- Часто белок рассматривается как жёсткое тело

Как химоинформатика может помочь?

- Разработка методов и управление информацией об 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.

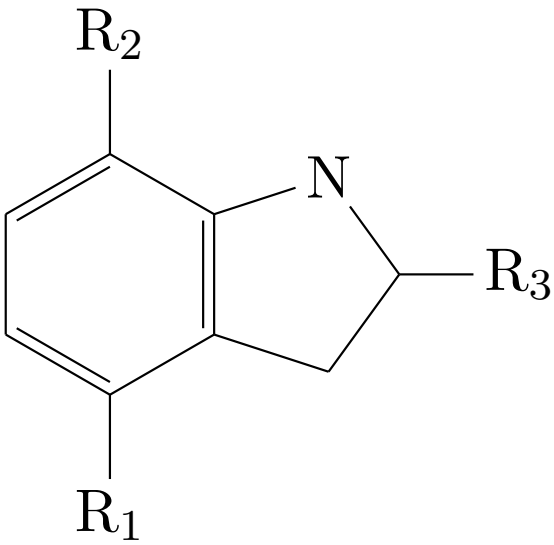
Пример, комбинаторная химия

- 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"



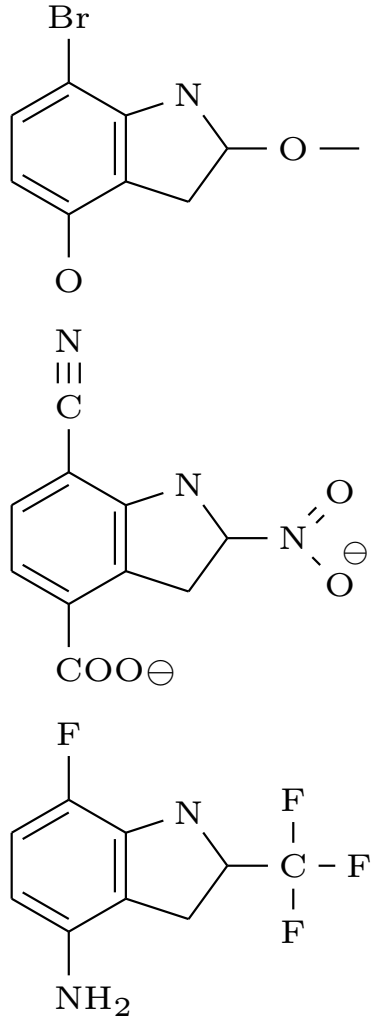
"Блоки"

R_1 = OH, OCH_3, NH_2, Cl, COOH

R_2 = Phe, OH, NH_2, Br, F, CN

R_3 = CF_3, NO_2, OCH_3, OH , PheO

Примеры



Химоинформатика и библиотеки

- Какие блоки выбрать?
- Какие библиотеки строить?
  - Дополнение известных наборов
  - Модификация под конкретный белок
  - Полное "насыщение"библиотеки
- Компютерное 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
    0.7145   -0.4125    0.0000 O   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
    0.0000    0.8250    0.0000 O   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

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

Реакции в виде SMILES

[I-].[Na+].C=CCBr >> [Na+].[Br-].C=CCI	реакция замещения
(C(=O)O).(OCC)>>(C(=O)OCC).(O)	образование сложного эфира

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
OCC	CCO
[CH3][CH2][OH]	CCO
C-C-O	CCO
C(O)C	CCO
OC(=O)C(Br)(Cl)N	NC(Cl)(Br)C(=O)O
ClC(Br)(N)C(=O)O	NC(Cl)(Br)C(=O)O
O=C(O)C(N)(Br)Cl	NC(Cl)(Br)C(=O)O

Описание SMILES: атомы

- Однобуквенные 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+]

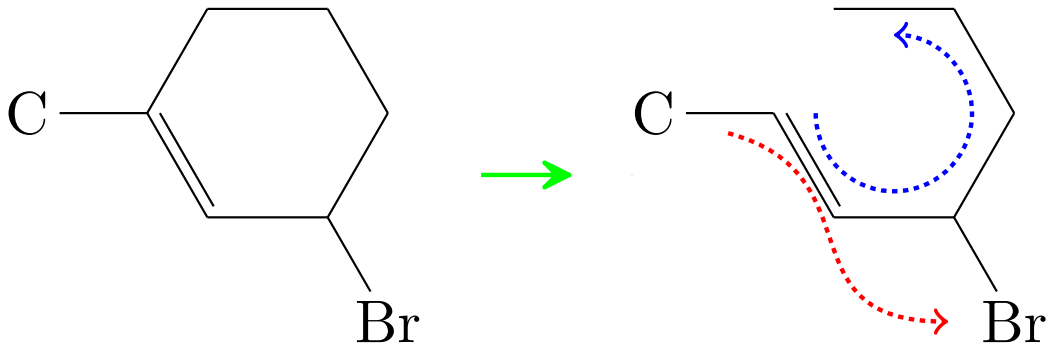
Описание SMILES: связи

CC	этан
C=C	этилен
O=C=O	CO2
C#N	HCN
CCO	этанол
[H][H]	водород

Ветвление цепи отображается в скобках ()
Пример: CCC(CC)COO

Описание SMILES: циклы

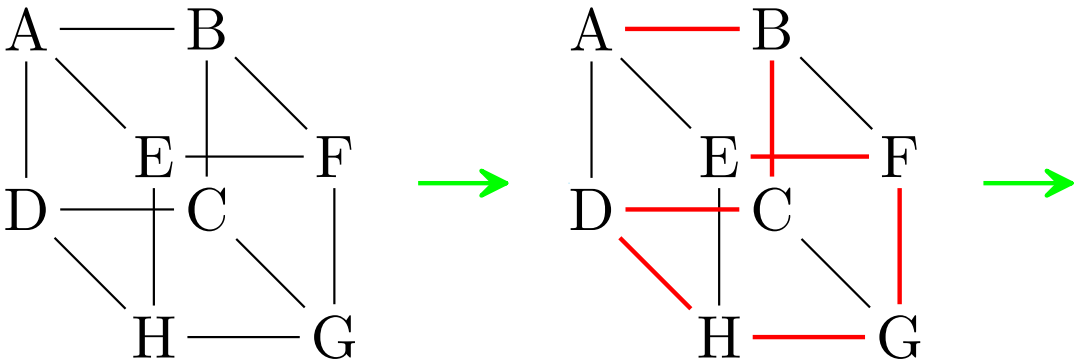
- C1CCCCC1 циклогексан



a) CC1=CC(Br)CCC1

b) CC1=CC(CCC1)Br

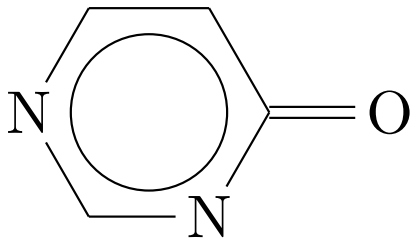
- Или более сложный пример:



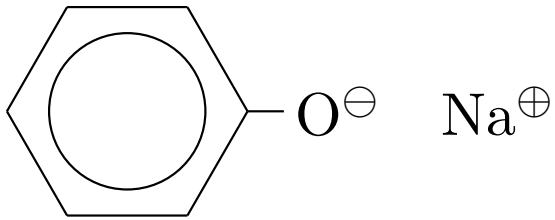
A14B2C3D1H5G3F2E45

Описание SMILES: ароматика

- SMILES для определения ароматичности использует расширенный алгоритм Хюккеля.
- c1ccccc1 eq C1=CC=CC=C1 тут все атомы находятся в sp2-гибридизации
- c1cccc1 eq C1=CC=CC1 , последний атом в гибридизации sp3.
- Ароматичными могут быть атомы: C, N, O, P, S, As, Se, и *.
- Пример: c1cnc[nH]c(=O)1



Структуры где есть нековалентные связи

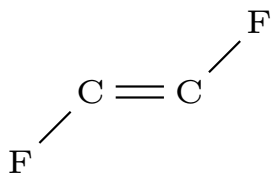


В SMILES нотации это:
[Na+].[O-]c1ccccc1
или
c1cc([O-].[Na+])ccc1

Изомеры

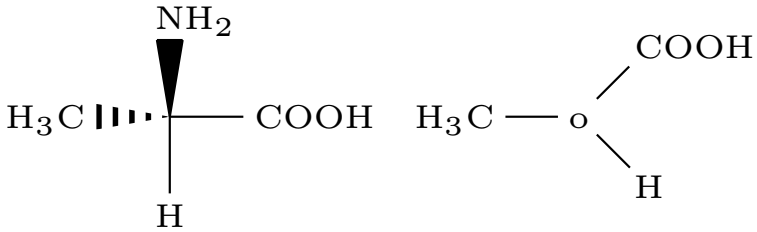
Изотопы --- [12C],[13C]

Цис-Транс



F/C=F/C или F\C=F\C

Хиральность



N[C@](C)(C(=O)0)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 водородами)
[X3]	атом с тремя связями, включая водороды
[v3]	атом с валентностью 3.

SMARTS: логические операторы и примеры

Логика:

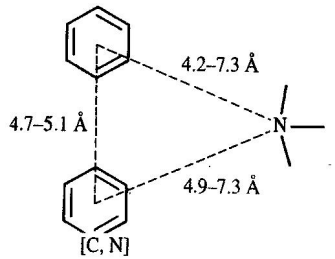
!e1	not e1
e1& e2	a1 and e2
e1,e2	e1 or e2
e1;e2	a1 and e2

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

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

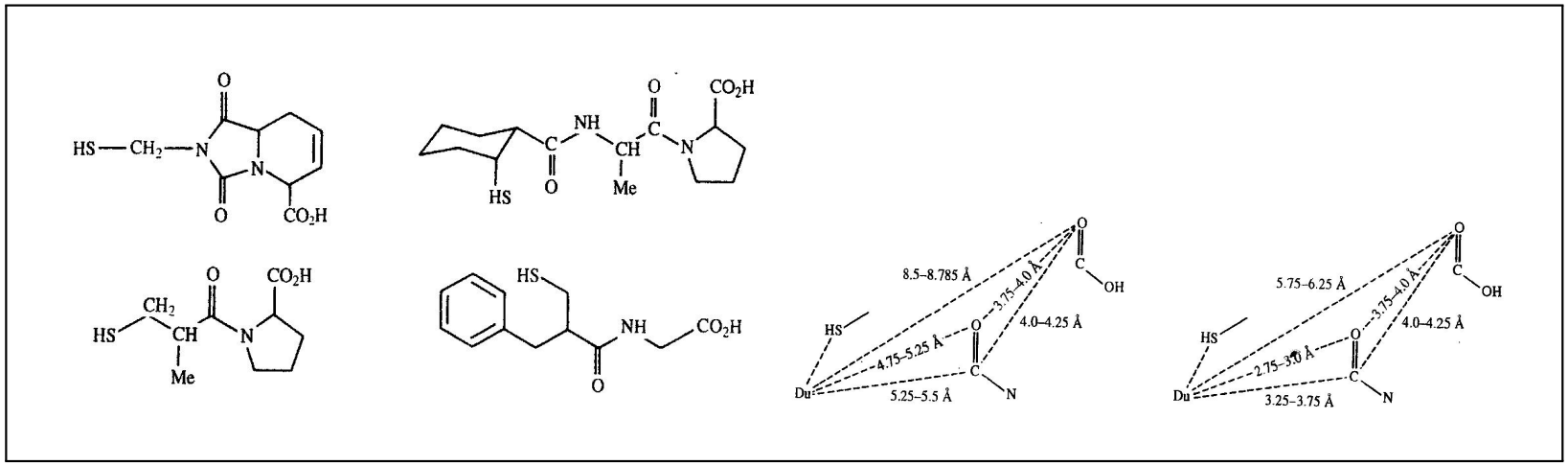


Проблемы с фармакофорами

- 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

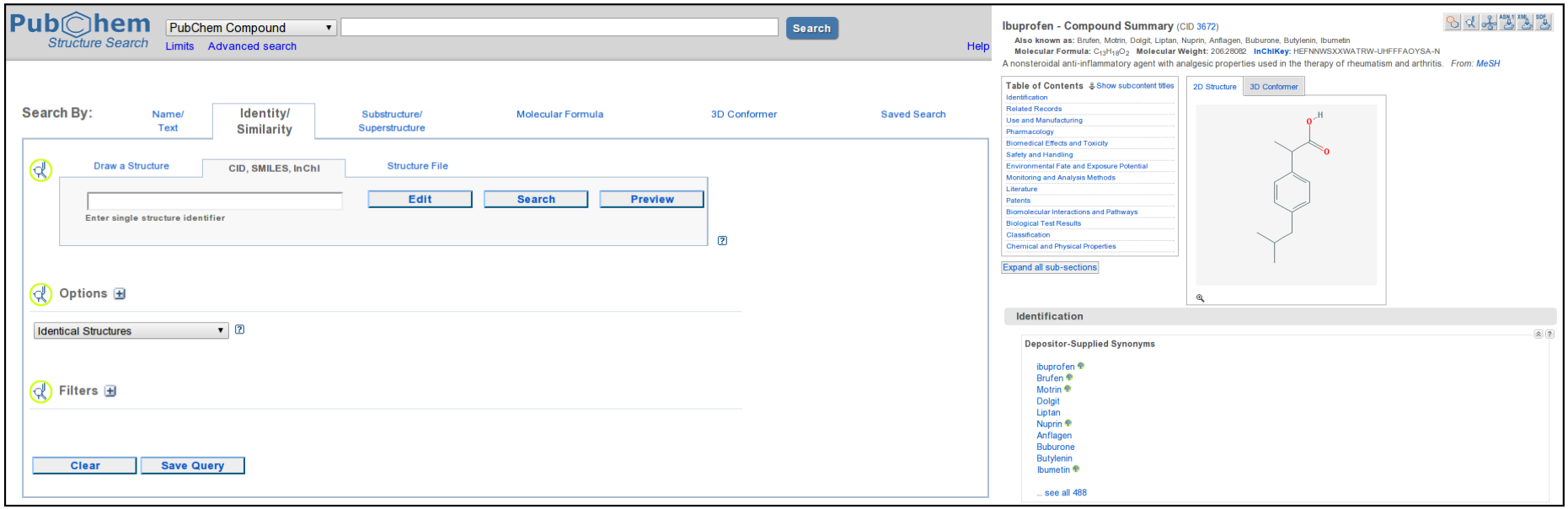
- Есть проблема:



- 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



SOAP доступ к PubChem

```
use SOAP::Lite; # +trace => qw(debug);
import SOAP::Data qw(name);

# Create PUG SOAP service object
my $pug_soap = new SOAP::Lite
    uri => "http://pubchem.ncbi.nlm.nih.gov/",
    proxy => "http://pubchem.ncbi.nlm.nih.gov/pug_soap/pug_soap.cgi";

..........
..........
$smile= $ARGV[0];

my $StrKey = InputStructure($smile, "eFormat_SMILES");

# IdentitySearch
my $ListKey = StructureSearch($StrKey, 250000);
print GetListItemsCount($ListKey) . "structures␣found\n";

my $url = Download($ListKey, "eFormat_SMILES", "eCompress_GZip");

print "Downloading␣...";
system ("wget␣-O␣$name\_smiles.gz␣$url");
print "Done!\n";
0;
```