自动化网络药理论文构思

## 论文修改构思

### 药物成分获取模块

1.多元数据模型的数据分析与处理，去重的处理，筛选的算法,OB,DL数据，基于ADME参数的筛选。

2.需要设计2个数据库表。处理药物名称、成分名称、OB、DL、PubChemdID等表的设计。

### 疾病靶点获取模块

1.基于启发式算法的基因靶点获取，基因获取的时候存在着靶点蛋白数量很多，需要进行进一步的筛选。使用

2.需要设计2个数据库，涉及到靶点的Gene名称、Protein name、UniprotID、PDBID等相关信息。

### 对接模块

1. 靶点对接对接算法，参考对应的文档进行
2. 可以解决SystemDock存在的不能批处理的问题，本文使用批处理的方式进行。本系统可以基于SystemDock的基础上进一步进行快速的全面的对接操作。

### 文件处理难点

1. 本文在进行药物成分处理的时候，针对非草药成分，可以让客户根据自己的查找直接上传查找结果，这里未来可以根据研究的进展进一步的进行研究，使用现成的框架，进行文本挖掘的方式进行药物成分的获取。
2. 本系统可以进行常见中药和复方药物的成分的分析，适用性广。
3. 本平台使用了很多的数据库，但是现阶段只是选择了其中部分数据库进行数据的筛选，后期可以进一步进行深入的数据库的准备，将更多的数据库加入到对应的研究中去。
4. 蛋白相互作用，这个是后续需要进行设计的亮点与关键之一。

## 论文概况

现在，根据实验的结果，本文可以将文章划分成5个部分，后续将按照下面的5个部分来写。

1. 综述，介绍先关的背景知识，然后介绍自己每个部分的内容。
2. 介绍网络药理学相关知识，同时介绍现在网络药理学研究中遇到的一些难点或者是挑战性问题。另外，介绍selenium+python的相关的知识与应用背景。
3. 介绍自己的框架，同时介绍自己在实现这个软件的时候，实现的每个部分的详细的描述，同时介绍每部分的功能与使用的相关技术。
4. 详细介绍自己每部分实现的情况以及网络药理的案例介绍。
5. 结论，介绍本实验的结论。

暂时的分析思路是按照上面的5个部分进行，其中会有一些细节数据需要简单的进行进一步的处理。

## 阶段化工作

这里根据前述的论文概况，进行了工作的简单再继续，在上述工作的基础上，进行了工作的细化安排。

首先，针对上述的第2个小模块的任务，这里提出了几个关键性的难点。这部分，首先需要阅读上次阅读到的PHD论文，分析这个论文中提到的一些挑战性问题。从这些问题中找到一些有价值的可以进行下一步工作的东西。

1. 数据获取。原始存在的异构、多源、规模大的问题
2. 选取靶点中遇到的问题。在选取靶点的时候，遇到一些例如靶点数据混乱，靶点数据的获取，不同的数据库有不同的效果等情形。这部分其实可以使用一些类似于机器学习、遗传算法的算法来求解。
3. 对接中存在的问题。在进行分子对接的时候存在的问题以及解决的思路。
4. 结果的分析。根据前一个阶段的结果，这里进行进一步的分析，主要是进行对结构的结果的分析。

第3个小模块也是可以进行一番的细化工作的，这里主要做的是一个云服务平台，其中关键的有这么几个部分，下面的几个小点是关键性的控制节点。

1. 给云服务平台取一个名字
2. 画出软件的架构图
3. 核心算法的设计

核心算法主要是进行整个流程中涉及到的一些算法的设计。这里举例的就是，

1. data analysis’s algorithm
2. 对接算法
3. 实验，case study。在进行实验的时候，需要给出 1.数据表 2.分析过程 3.结果可视化。

## 相关概念与关键

数据获取模块 数据获取算法

箭头，删掉， 靶点算法

靶点名字修改

文件到处的类型 对接算法 （该部分没法写），这部分可以引用SystemDock中的算法，对接部分可以直接引用。

框架设计:本框架提供基于MVC架构的在想对接平台，这里实际上使用的是MVC架构一个网站，这里会介绍该MVC架构的具体设计，架构的每一部分的设计。例如，前端平台的设计，后台的设计，

怎样的在线方式，现有的云平台技术，基于某种架构

现有的数据方法，自己提出的方法，

靶点数据多元异构，数据量特别大，提炼数据需要的是

靶点算法

(筛选) 根据人的特征的筛选，面向xxx的靶点蛋白的筛选(筛选算法,使用启发式算法来进行，例如遗传算法，模拟退火算法，粒子群算法，具体的思路是选择一个python的框架来做可以)，

在进行靶点蛋白筛选的时候，存在这样的问题，就是靶点蛋白的选择问题，这里在进行靶点蛋白筛选的时候，遇到的问题是，很多情况下，靶点数据很多，但是进行

对接算法的设计，对接算法参考SystemDock Web Site这个在线对接平台的设计，使用的算法就是这个平台对应的对接算法。

除了上面的页面之外，这里进行了前端页面的设计，前端页面的设计是怎么样子的，这个需要详细设计。

对接算法， systemsDock applies AutoDock VINA (28) to perform docking simulation.

## 后台结构设计

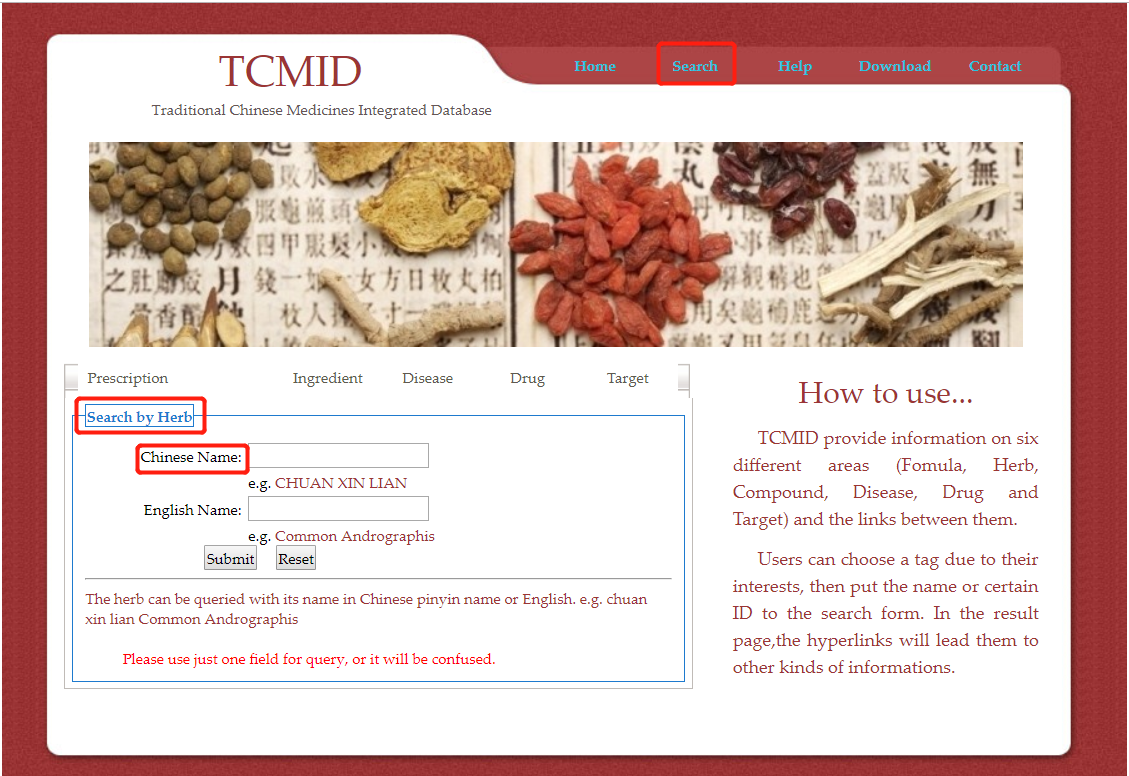
本项目按照图示设计了软件的框架，核心主要是分成三个部分，分别是化合物成分获取模块、靶点获取模块和对接对接模块。从结构图上可以看到，化合物模块设计了两个部分，分别是根据名称获取到药物的成分、根据成分名称下载对应的3D结构。靶点模块设计了两个部分，分别是根据疾病的名称获取到疾病的靶点基因名称、根据靶点基因名称获取对应的合适的靶点蛋白。对接模块，根据靶点蛋白和化合物成分进行对接。



化合物的成分模块分别是化合物成分获取和成分3D文件下载部分。化合物名称获取，本实验是从TCMSP和TCMID数据库两个数据库来获取到的成分数据。下面的介绍的成分获取模块划分为化合物成分名称获取，这个部分的数据主要是从下面的两个网站TCMID、ICMSP数据库中查找对应的药品化合物成分数据。首先，看到的是TCMID网站如下所示，

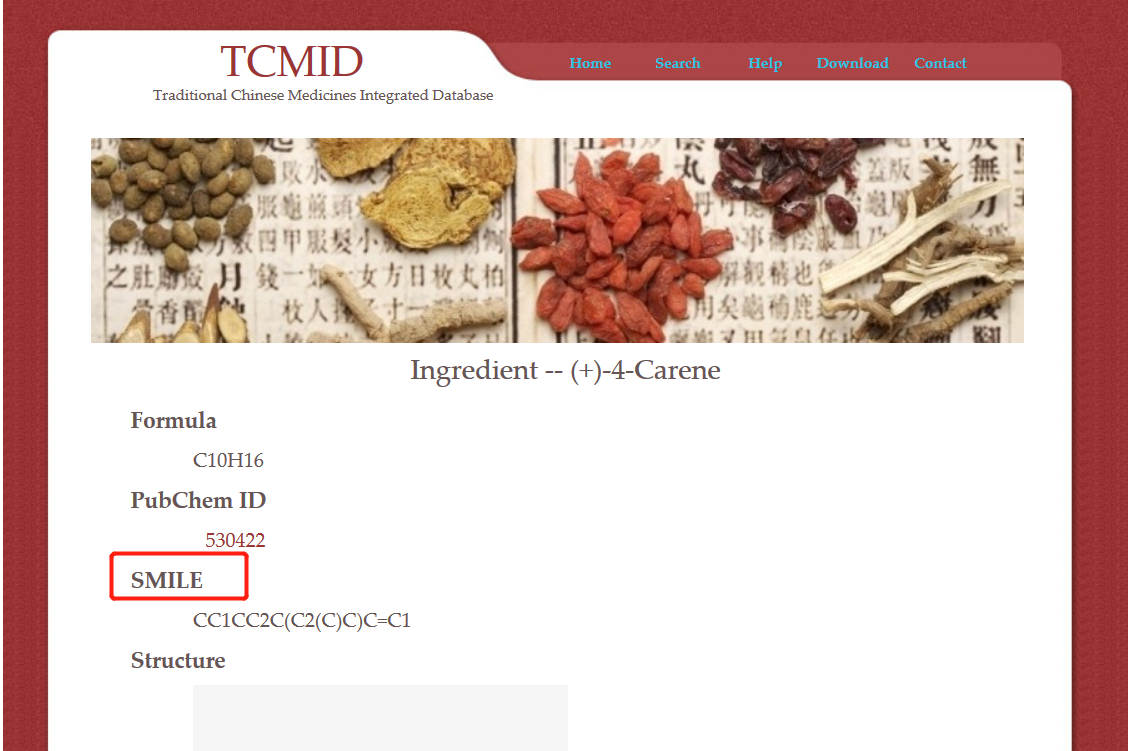


本实验需要查询的药品窗口如下所示，



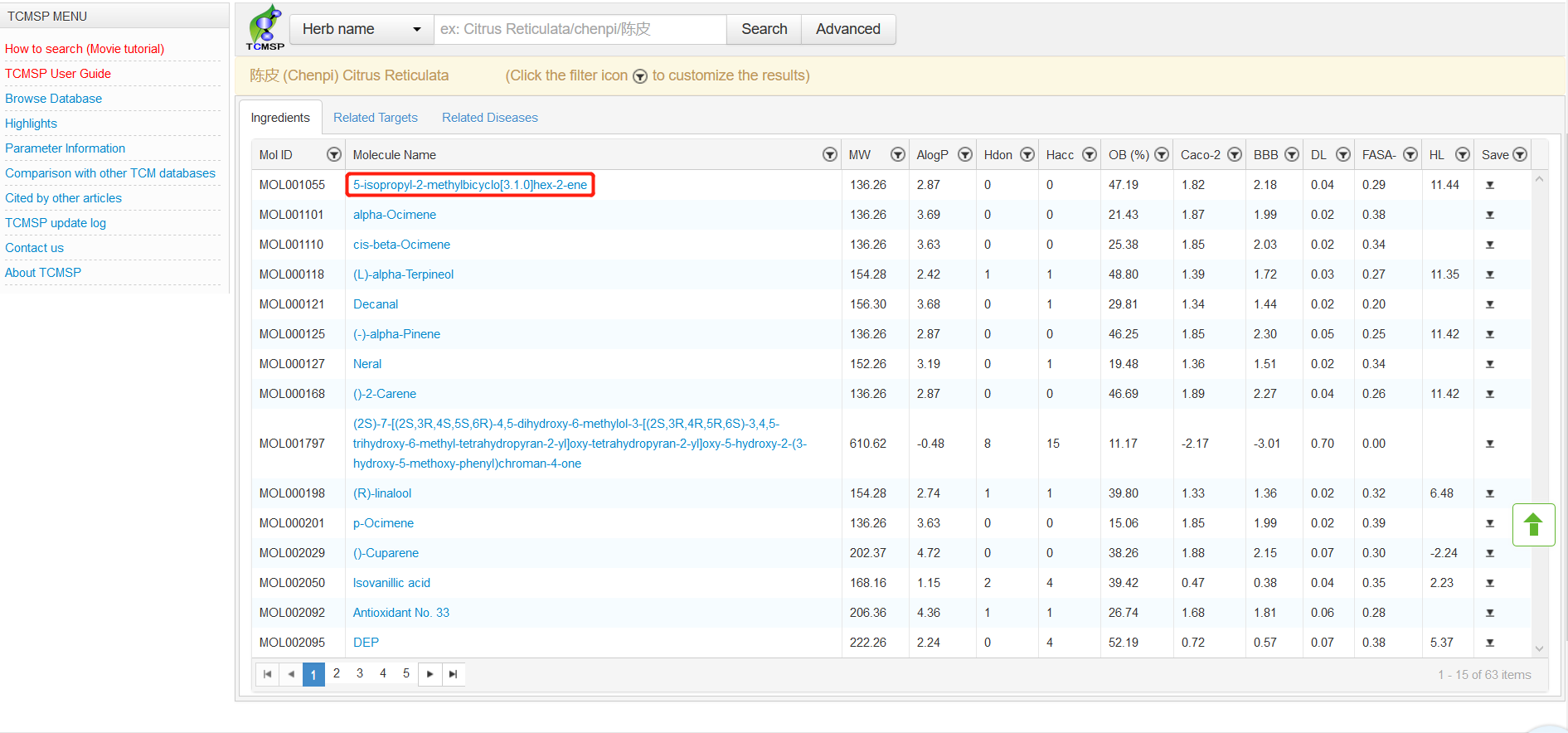
上述的窗口就是药品查找的窗口，这里的查询窗口进去后可以得到对应的药品成分的信息，例如这里给的ER SHEN WAN,这就是一个Herb,输入该信息后查询可以得到对应的结果页面，

可以查看到对应的成分信息，点击进去查看得到每种成分的具体信息，



这里有一些结构信息，因为本实验使用的是对应的药品成分的名字获取成分3D结构，所以这里这里选择的时候，关键信息是药品成分的名称。

成分信息获取的第二个数据库是TCMSP网站，下面是该网站的查询页面，下面的页面是查询药品成分的首页，输入查询的草药(Herb)名称，可以获取到对应的药品成分信息，例如这里数据“陈皮”,可以查询到对应的成分数据。



本项目根据上述的两个网站获取到对应的药品成分名称信息。

接下来是成分3D结构的下载，这里使用了PubMed数据库进行3D结构的下载，在PubMed数据库中，选择PubChem Compound，在输入框中输入化合物名称，可以查询得到化合物的结构等相关信息，因为某种名称的化合物对应的成分在不同的研究文献中可能会存在些许的差别，但实际上都是同样的成分，选择一个，可以查看到对应的3D结构，然后将这个3D结构下载下来即可。

|  |
| --- |
|  |
| PubMed化合物成分查询页面 |

下面是化合物信息页面，对应的3D Conformer就是成分的3D结构信息，这是关键数据。

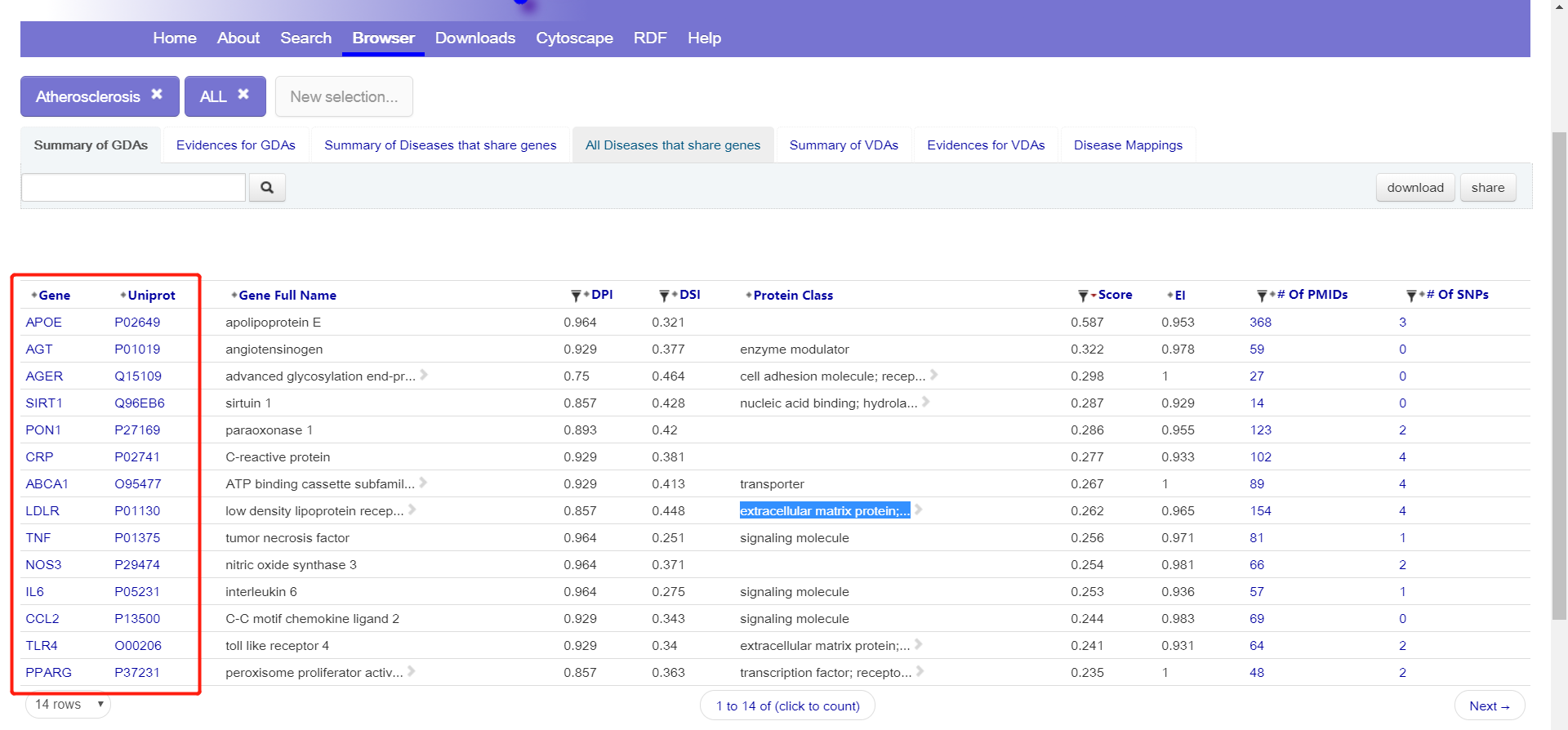
|  |
| --- |
|  |
| 化合物成分信息 |

在上面的药品成分的获取和成分3D结构的下载是成分获取模块关键的部分。通过对这部建模，实现自动化操作完成原始药品成分的筛选。

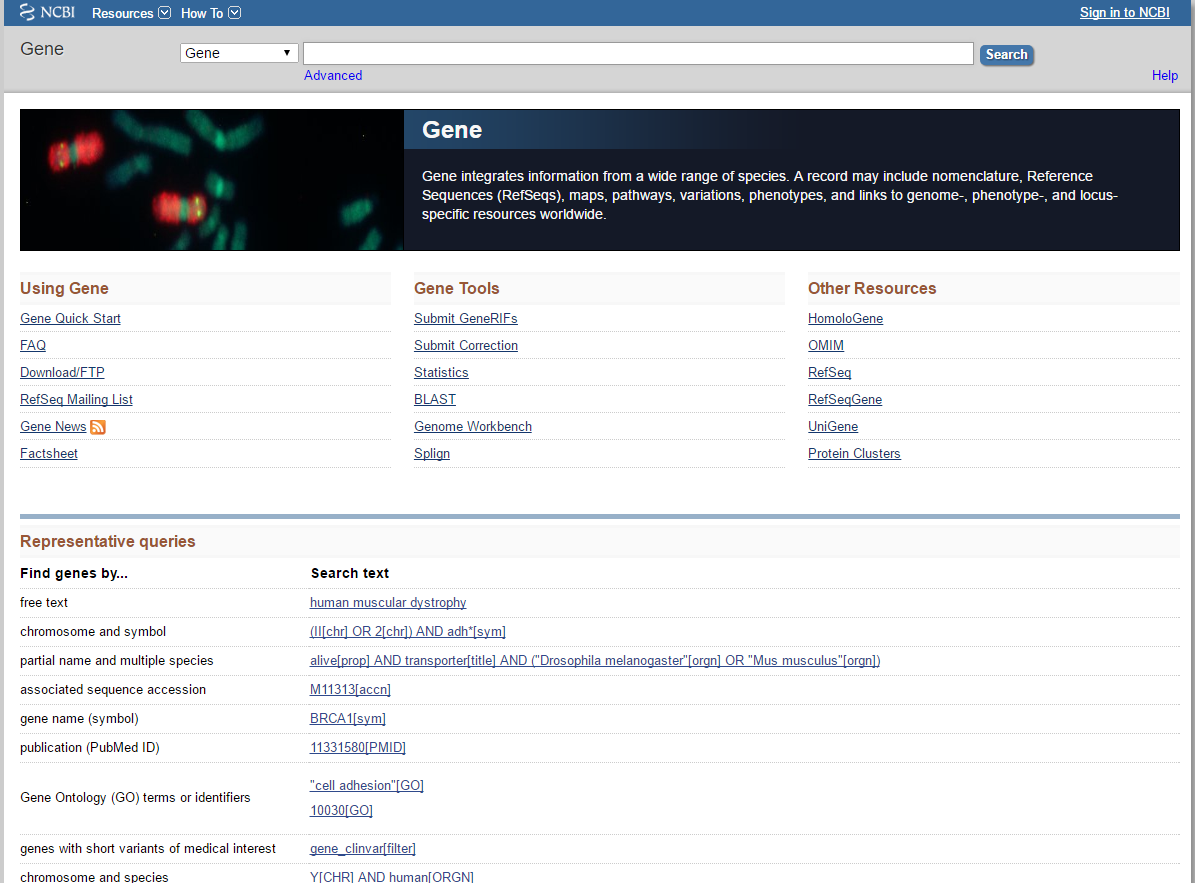
靶点获取模块主要是获取疾病对应的基因靶点和靶点蛋白的获取两个方面。这里使用了

DisGeNET、NCBI、UniProt等网站数据库来获取靶点信息。

首先，对于一般的疾病来说，在数据库DisGeNET中可以直接查询到这个对应的靶点信息，下面就是这些靶点的样例，这里根据数据可以找到对应的很多靶点



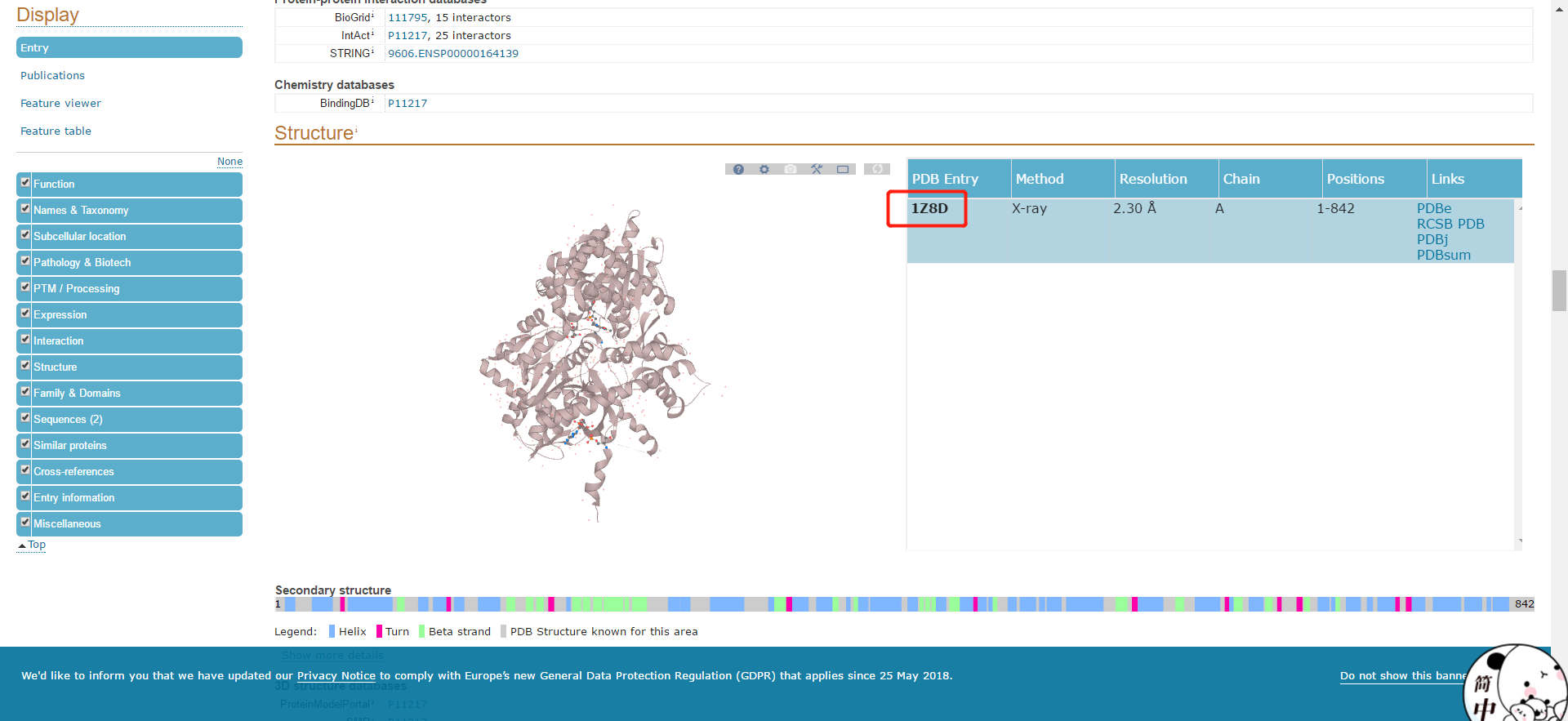
这里的靶点数据可以直接拿来进行使用。但是有些数据是不能直接找到靶点的，这就需要另外的数据库来处理，这里根据NCBI数据库进行查找，下图是NCBI数据库页面，

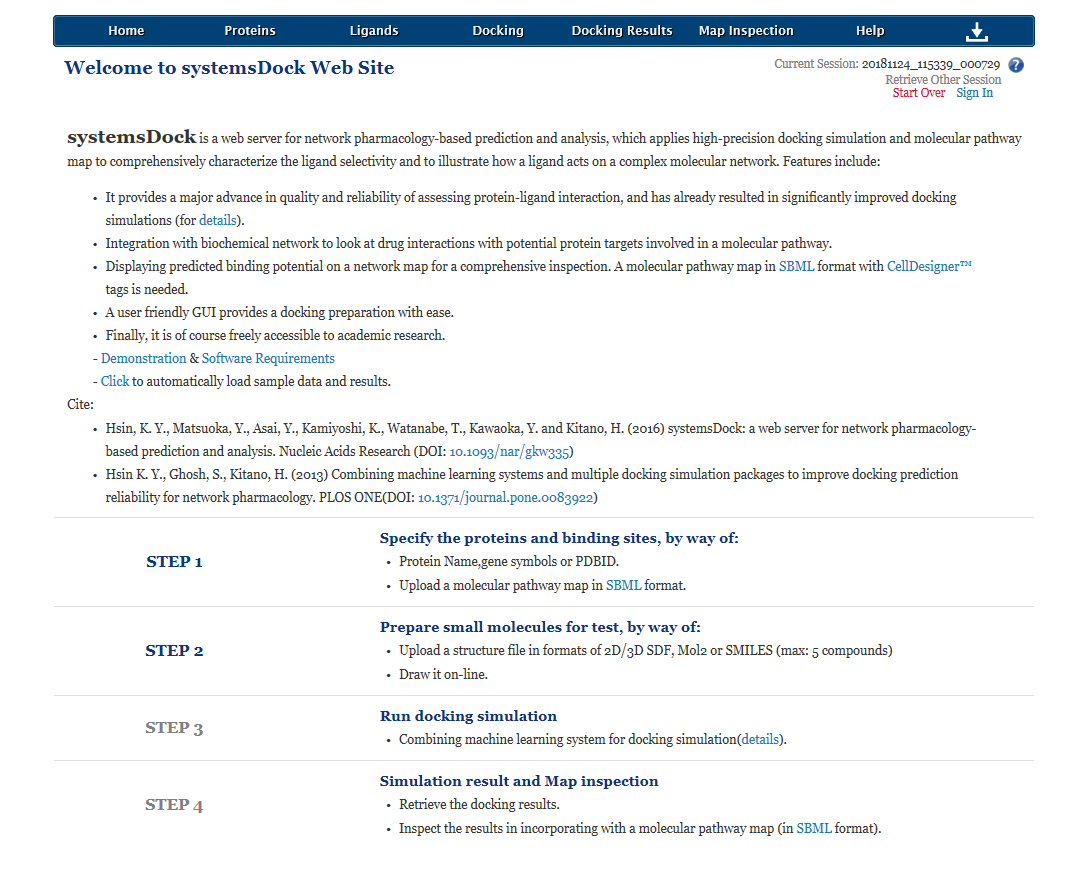


下面从这个数据库里查询靶点信息，下面的页面中输入的疾病名称，然后筛选出人类基因数据，就可以获取到对应的靶点信息，靶点名称如下所示，但是这个数据其实是不能直接使用的，因为对接使用的是靶点蛋白信息，所以需要在基因靶点的信息的基础上继续获取靶点蛋白数据，这里先将基因靶点数据遍历存储，然后再在UniProt靶点蛋白获取页面遍历筛选出所有的靶点信息。靶点信息页面会得到靶点蛋白的PDB ID编号，由于一种基因蛋白对应的靶点蛋白存在多个靶点蛋白，这里根据一定的选择原则，对靶点蛋白进行筛选，筛选后的数据存入到对应的数据结构中。另外，使用DisGeNET数据库获取到的靶点基因对应的UniProt

|  |
| --- |
|  |
| NCBI基因靶点获取模块 |

|  |
| --- |
|  |
| UniProt靶点蛋白获取页面 |

的id信息，这个可以在进行靶点筛选的时候直接跳过根据基因靶点名称的初选，从而更方便的进行下一步的选择。

至此，完成了化合物成分和疾病靶点的选择，下一步是对接操作，对接使用的是SystemDock软件，该软件是在线的对接系统，底层使用的AutoDock Vina软件，下面介绍这个软件的使用。该软件总共分为4个模块，分别是蛋白质模块、配体、对接、对接结

果。

按照对接的流程，分别是将前述获取到的靶点蛋白数据上传到对应的位置，靶点蛋白上



传使用的是靶点蛋白的PDB ID，接下来是配体的页面，把下载下来的配体上传到页面的对应位置，接下来是配体的上传，配体的上传是上传的配体的3D结构，最后进行对接操作，对接结果会传到对应的邮箱。在上述的操作完成后，可以查看到对接结果可视化的页面。从页面中可以查看到不同的靶点蛋白和化合物成分的对接打分。

|  |
| --- |
|  |
| 化合物成分上传 |

|  |
| --- |
|  |
| 对接结果可视化 |

### 数据库设计

这里根据前面的数据分析进行了数据库的设计，数据库涉及到药物成分数据库以及靶点数据库的设计。药物成分数据库涉及到药材成分表、成分详情表两张表，下面的第一张表是成分详情表Compound。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键id | int(11) | 11 | PRI |
| Molecule | 分子名称 | varchar(256) | 256 |  |
| OB | OB | varchar(32) | 32 |  |
| DL | DL | varchar(32) | 32 |  |
| PubChemCID | PubChemCID | Varchar(32) | 32 |  |
| drug | 药品名称 | Varchar(32) | 32 |  |

建表语句：

CREATE TABLE Compound(id int(11) primary key auto\_increment, molcule varchar(256),OB varchar(32), DL varchar(32), PubChemCID Varchar(32));

接下来药品成分表drugs,该表需要设计3个字段，分别是id、drug、module，用于记录每种药物的成分信息。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键id | int(11) | 11 | PRI |
| drug | 药品名称 | Varchar(32) | 32 |  |
| molcule | 分子名称 | varchar(256) | 256 |  |

建表语句：

CREATE TABLE drugs (id int(11) primary key auto\_increment, drug varchar(32),molcule varchar(256));

数据处理完成后数据会存储到对应的数据表中。

靶点数据库的设计如下所示Target：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键id | int(11) | 11 | PRI |
| Protein | 蛋白质名称 | varchar(256) | 256 |  |
| Gene | 基因名称 | varchar(32) | 32 |  |
| UniProtID | UniProt ID | varchar(32) | 32 |  |
| PDBID | PDB ID | varchar(32) | 32 |  |

上述的信息存储的是靶点的数据，但是还有一个基础的数据，需要存储的是疾病和基因之间关系的数据，该表名称为diseases。这张表的字段有三个，分别是id、disease、gene用于存储基因的靶点等相关信息。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键id | int(11) | 11 | PRI |
| illness | 疾病 | Varchar（32） | 32 |  |
| gene | 基因靶点 | Varchar（32） | 32 |  |

对接完成后，需要对后续的数据进行处理，得到对接的表docked,该表有几个字段，分别是PDBID、PubChemCID、score等

表设计docked

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键ID | int(11) | 11 | PRI |
| Origianl | 原始蛋白名称 | varchar(32) | 32 |  |
| Specified | 用户定义蛋白名称 | varchar(32) | 32 |  |
| PDBID | PDBID | varchar(32) | 32 |  |
| PubChemCID | PubChemCID | varchar(32) | 32 |  |
| Score | 对接打分 | varchar (32) | 32 |  |

建表语句：

CREATE TABLE docked(id int(11) primary key auto\_increment, Origianl varchar(32), Specified varchar(32), PDBID varchar(32), PubChemCID varchar(32),Score varchar(32));

上述的表完成后，需要进行多表关联，获取到最后的全部的信息，主要是靶点、PDBID、对接打分。

前面的数据库设计完成后，会给出最终的对接结果，在这些对接结果的基础上进行下一步的操作，这其实是可以行的。多表关联，获取到对应的最终数据，最终数据展示的应该如下所示：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Herbs | Molecule | OB(%) | DL | PubChem | gene | uniprotid | protein | PDBID | score |
|  |  |  |  |  |  |  |  |  |  |

上述就是进行后期进行数据分析的时候关键的表，其中的herbs等数据是来自于成分表，后面的gene等数据来自靶点表，他们之间使用对接结果表进行关联，三张表进行关联可以得到最终的表。

现在遇到的问题是数据不准确的问题，数据不准确的问题是一个比较大的问题，这个问题在进行数据处理的时候是比较严重的问题，这里处理得到的问题主要是进行的顺序存在问题，现阶段需要做的应该是完全设计好分子数据库和靶点数据库，在

## 前端结构设计

本实验还涉及到前端的设计，前端的设计的整体流程，前端的整体流程首先是分别完成成分获取、靶点获取、分子对接等几个模块。下面是简单的框架，这里甚至给出了一个关键的



前端设计的时候设计几个框架，分笔试

## 具体实现

上述的对接整体流程的和核心是基于python+selenium进行设计的，

## 基本情况

本文基于自动化技术进行网络药理学的基础数据的处理，并最终可视化实现网络药理学药品。这个的实现过程需要一些详细的细节，分别是小分子化合物的获取，小分子化合物的获取也有两个类型。分别是使用SMILES或者是使用ChemDraw来进行绘制。但是核心都需要获取到中药的成分信息。

案例的选择，本文实际上并不需要具体在特定的某种药物上选择数据进行筛选，仅仅完成前期的数据的获取和筛选就好。本文进行的相关实验，可以方便后续网络药理学相关的研究，大大推进实验的进展。

## 文章构思

1.算法设计思路，数据分析软件模块构建。

2.特征，创新点，本部分重点考虑可信数据建模，也就是可信数据模型。

3.体现的核心实现技术。也就是数据分析数据挖掘技术。

## 软件算法

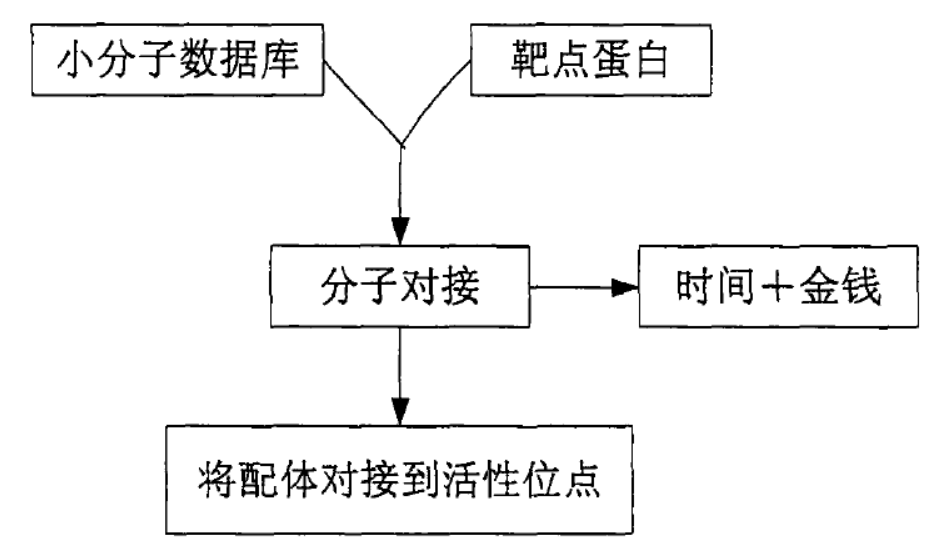
现在选择的数据库主要是两个数据库，TCMSP和TCMID这两个数据库，现在分别就这两个数据库进行相关的算法设计。软件设计的主要目的是为了准备小分子配体的结构信息而进行的相关的操作。整体思路是，使用selenium+python的技术组合搭建自动化环境，进行药物分子数据的获取。在具体的技术实现上，需要使用网页分析技术，分析具体页面的特点，使用selenium模拟人的点击操作，最终获取到对应的数据，使用python的DataFrame数据进行数据的存储，并最终将数据到处到。

### 技术流程

本软件实现的是从TCMSP/TCMID/上海有机所三个数据库的数据找到相关的分子的数据，靶点数据从drugbank数据库获取到。然后按照下面的流程，实现整个系统的数据走向。从数据流图可以看出，该实验只要是将药品的成分和靶点蛋白质的获取划分为两个模块，分别从对应的数据库获取。将分别获取到的小分子化合物和靶点蛋白存储下来作为下一步对接的基础。其中，这里提及到的discovery studio和最终实现的网络可视化这里不做细节分析。



除了上面的数据网络流图之外，这里还有一部分是整个对接过程的流程图:



这部分的

前述相关的设计是数据的抽取部分，现在还能做的就是ADME算法，这部分主要做的是进行数据的筛选操作，筛选工作也可以做一些自动化实现。

### TCMSP设计

对于从TCMSP中获取到需要的生物小分子数据，主要根据药品的名称信息，获取到对应的小分子的成分数据。例如，这里的程序可以实现的是，获取到小分子ID(Molecule ID)，分子名称(Molecule name),分子图片，分子图片地址等信息。下面是程序的构造:

获取药品成分的所有url信息

解析查询页面，获取到成分信息

下载所有化合物成分的图片

下面会介绍上述的三个组成部分的算法

获取药品成分的所有url信息

输入数据：药品名称

给定TCMSP首页URL并访问

定位输入信息框，并清空信息

定位输入信息框，输入药品名称

定位搜索按钮，点击确定，页面发生跳转

创建新数组hrefs用于存储药品成分的页面url

定位首页的表格信息，获取首页的表格中药品成分的URL信息

将这些url存储到hrefs中

定位获取到总共的页面数目n

从1到n循环：

定位当前页面的表格信息，获取首页的表格中药品成分的URL信息

将URL数据存储到hrefs中保存

返回hrefs

解析查询页面，获取到成分信息

输入数据：成分页面的url

解析页面的url

新建空数组info,存储成分信息

分别获取成分的Molecule ID，Molecule name和imgs信息

将上述获取到的信息存储到info中

返回info

下载所有化合物的图片

输入参数：存储图片的文件夹地址，化合物信息数组

判断地址文件夹是否存在

If 是

PASS

Else

新建文件夹

For 化合物信息数组中的每个化合物信息

取出每个化合物信息中的url信息

分别将该url对应的图片解析下载存储在本地文件夹中

完成后输出“抓取完成”信息

### TCMID设计

另外一个重点的数据的获取是通过TCMID数据库挖掘到对应的相关信息。该模块的设计同样需要涉及到获取药品成分的名称、化学式、SMILES格式、图片等模块。每个模块在程序中是可以单独的存在的。这里获取到的SMILES并不是全部都存在，可能仅有一部分是存在，因此，在实验中，会将不存在的那些使用空值来补足。

TCMID软件的模块设计

获取到页面的药品链接信息

找到对应药物的页面

解析页面，获取到需要的成分信息

下载成分图片

接下来是每个模块的实现算法，

获取到页面的药品链接信息

输入参数：页面URL

解析页面

创建一个空数组urls用来存储

找到页面中的表信息

分别获取表中的各项成分url信息

并将信息存储在urls中

返回urls

解析页面，获取到需要的成分信息

输入参数：页面URL

解析页面

分别获取名称、化学式、id、SMILES、结构地址

创建空数组info，分别将上述信息按顺序存储在数组中

返回info

找到对应药物的页面

输入参数：药品名称

给定主页url

访问主页

定位查询窗口，模拟点击操作

定位信息输入窗口，清空输入框

定位信息输入窗口，模拟输入药品名称

定位信息输入窗口，模拟键盘Enter操作

窗口转换到操作完后的的页面

获取转换后的页面url1

返回url1

下载化合物的图片

输入参数：存储图片的文件夹地址，化合物信息数组

判断地址文件夹是否存在

If 是

PASS

Else

新建文件夹

For 化合物信息数组中的每个化合物信息

取出每个化合物信息中的url信息

分别将该url对应的图片解析下载存储在本地文件夹中

完成后输出“抓取完成”信息

### 上海有机所成分库

上海有机所的成分库也是获取小分子数据的一个重要的来源，对该数据库的数据的爬取也是一个重要的数据来源。为了获取到小分子的结构信息。该部分需要处理的主要是需要获取到药品的小分子成分，小分子的SMILES格式，以及各成分的图片信息。根据上述的分析，本文进行了

程序的模块划分：爬虫模块划分为几个模块，分别是实现网页的登录，模拟访问网页，然后登录进去后，选择访问的药品，然后从药品中全名筛选出全名匹配的药品，最后再根据找到的药品的名称进行选择。

这里，同样需要对数据的情况进行分析。这里根据查询的结果划分为三类，第一类是没有结果，第二类是结果只有一种情况，第三类是结果有很多种，具体来说，就是查询“三七”，这里的结果不仅仅是“三七”,而是其他的一些，例如“海三七”等其他的一些药品。这种情况下，本实验室需要对脚本进行调整吗，需要筛选出特定的“三七”才是合适合理的。

基于上海有机所的成分数据库获取药品的成分的程序由一个主要的程序组成。

程序的实现算法如下，

给定上海有机所查询中药的首页地址并访问

输入信息：药品名称、文件名称

定位用户名字，清空数据

定位用户名字，输入用户名

定位登录密码，清空数据

定位登录密码，输入用户密码

定位登录按钮，点击登录

定位中药药材检索链接，点击，页面会跳转

定位药材名称输入框，清空数据

定位药材名称输入框，输入药材名称

定位数据提交按钮，模拟点击

定位药草的“药材描述”链接，点击跳转到对应的页面

定位药材的“化学成分”链接，点击跳转到化学成分信息页面

定位获取到总共的页面数目n

创建新数组drugs来存储成分信息

For I 从1到n

分别获取化合物化合物名称、化合物分子式、CAS号

将上述的信息组成数组后再存储在drugs中

将drug信息写入到文件中

### SMILES数据获取

Pubmed数据库是获取到小分子信息比较关键的数据库，从该数据库上可以获取到对应的小分子的SMILES信息，该信息的获取分为两个模块，分别是页面查询模块和信息获取模块。另外，还需要设计批量数据获取程序。获取到对应的信息后，SMILES格式的信息与原始的分子名称信息应该存储在一起，然后保存在DataFrame格式，最后保存在Excel文件或者是MySQL数据库中。

再Pubmed查询数据中，有很多的成分数据，这里选取所需要的第一个小分子化合物的药品作为我们需要的分子信息。

在输入框中输入成分的名称信息后，这里需要注意的就是，需要分三种情况，第一种情况是，输入的数据没有结果，也就是说，把给定的小分子成分的名称输入到查询框的时候，没有结果；第二种情况是，输入数据后，仅有一种结果结果，导致会直接跳转到对应的查询页面，这种情况下假如使用原始假设的脚本，直接查询的话会导致页面定位错误；第三种情况是，输入数据后，会产生很多的结果，这种情况下，需要进行筛选，从中找出合适的小分子成分，这种情况下，本文的处理方式可以是使用自动化定位技术，确定第一个位置上的结果作为我们的研究结果。综合这里的第二条和第三条，合理的处理办法是，首先判断第三条，如果出现异常再判断第二条。

通过相关的分析与设计，本程序可以实现以下几个功能，分别是：

1. 给定excel存储化合物小分子的名称集合，可以将数据处理成list格式，方便批量获取SMILES格式。
2. 可以在输入给定的小分子的名称后，可以自动获取到该分子的SMILES信息，并可以将这个数据输出。
3. 可以给定批量的excel小分子的名称，使用自动化程序，获取到每条小分子对应的SMILES信息。

下面是这个程序的算法设计:

在查询分子界面上,

访问地址页面

定位SMILES

定位SMILES对应的文本信息并抓取

返回SMILES信息

在查询首页上，

访问查询首页

定位首页查询框，并清空查询窗口

定位首页查询框，并输入查询关键字

定位首页查询按钮，并点击确定

上述是分步实现查询与获取SMILES格式的步骤，下面将这两部分的伪代码结合起来。

访问查询首页

定位首页查询框，并清空查询窗口

定位首页查询框，并输入查询关键字

定位首页查询按钮，并点击确定。

上述操作完成后，页面会跳转到查询结果页面。分三种情况

If 查询结果只有一种

Then 直接定位SMILES

定位SMILES对应的文本信息并抓取

Else if 查询结果有多种

Then 定位查询结果框，选择第一个查询结果，获取链接属性

将链接属性作为结果页面的url

访问结果页面的url

定位SMILES

定位SMILES对应的文本信息

Else 查询结果为空

SMILES结果赋值为空

返回SMILES

上述的是单挑SMILES获取的算法。下面是自动化批量处理获取SMILES的算法

算法的输入是一个链表，链表的每个元素都是成分的名称

访问查询首页

给空数组Drugs存储查询结果

对SMILES链表中的每一个原始,使用循环操作:

定位首页查询框，并清空查询窗口

定位首页查询框，并输入查询关键字

定位首页查询按钮，并点击确定。

上述操作完成后，页面会跳转到查询结果页面。分三种情况

If 查询结果只有一种

Then 直接定位SMILES

定位SMILES对应的文本信息并抓取

Else if 查询结果有多种

Then 定位查询结果框，选择第一个查询结果，获取链接属性

将链接属性作为结果页面的url

访问结果页面的url

定位SMILES

定位SMILES对应的文本信息

Else 查询结果为空

SMILES结果赋值为空

把SMILES赋在Drugs后

重新将url更新为查询首页

访问查询首页,将页面调整到查询首页

返回Drugs

这个算法可以实现自动化查询的批量处理。

### 数据合并

现在的情况是，为了合并所有成分，例如从TCMID、TCMSP、上海有机所的数据，关键的点是PubChemID这个数据，现在的处理思路是TCMID的页面获取到对应的PubChemID

### 化学成分3D结果获取

现在计划使用pubchems数据库获取到给定的化学成分的3D结构，处理的思路有两个，分别是给定药品的名称，两步获取到对应的成分的3D结构下载，或者是先期获取到化学成分对应的页面的url，然后直接在特定的url上解析，下载对应的化学成分的3D结构。注意，这里使用的浏览器是Chrome浏览器，该浏览器的好处是下载的时候直接点击就会下载到对应的文件夹下。

现在给出最后一种方案的算法。

输入：成分页面的url

通过id定位到’ 3D-Conformer’

在上面的基础上通过class名称定位’menu-btn’取第一个元素并模拟点击

通过link\_text为‘Save’定位并点击下载操作

执行结果：文件被下载到本地

这个算法是给定页面的url在进行访问的方法，是比较好的，一是可以进行方便的操作，防止因为某种药品的成分不存在对应的url而导致的错误，二是可以进行对应的批量操作。

这里，需要做的就是根据药品的名称得到url，最后利用url来进行3D-Conformer的下载。

这里存在一个问题，那就是下载的3D的名称并不会存在

### 靶点数据模块

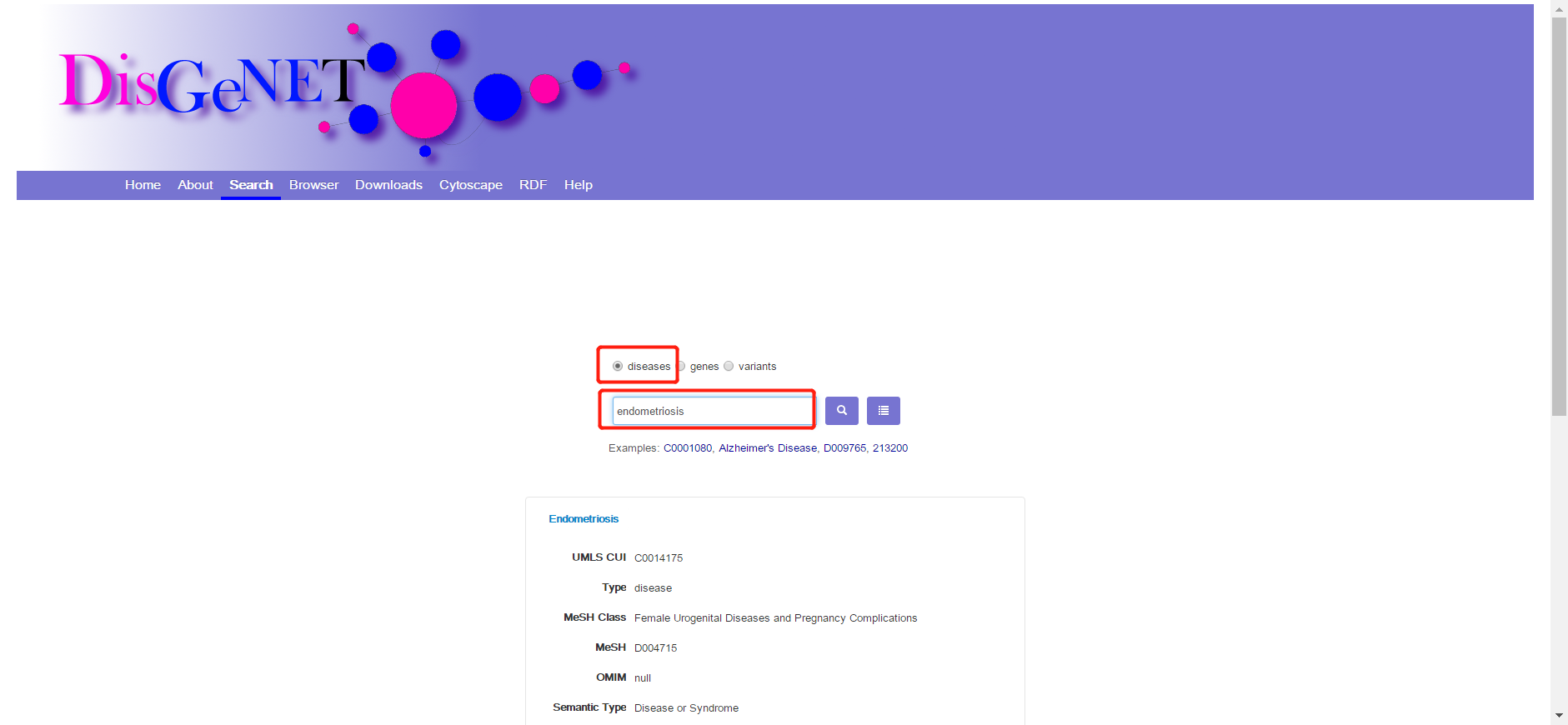
靶点信息的获取，关键的数据库是DrugBank数据库，从改数据库中查询到靶点信息，并将该靶点信息存储成DataFrame格式，最终存储到excel或者是MySQL数据库中，供后续的研究需要。

这部分靶点数据的获取实际上需要从蛋白质数据库(Protein Data Bank,PDB)筛选，找出对应的靶点数据，并且对这些靶点数据采用相关的实验。

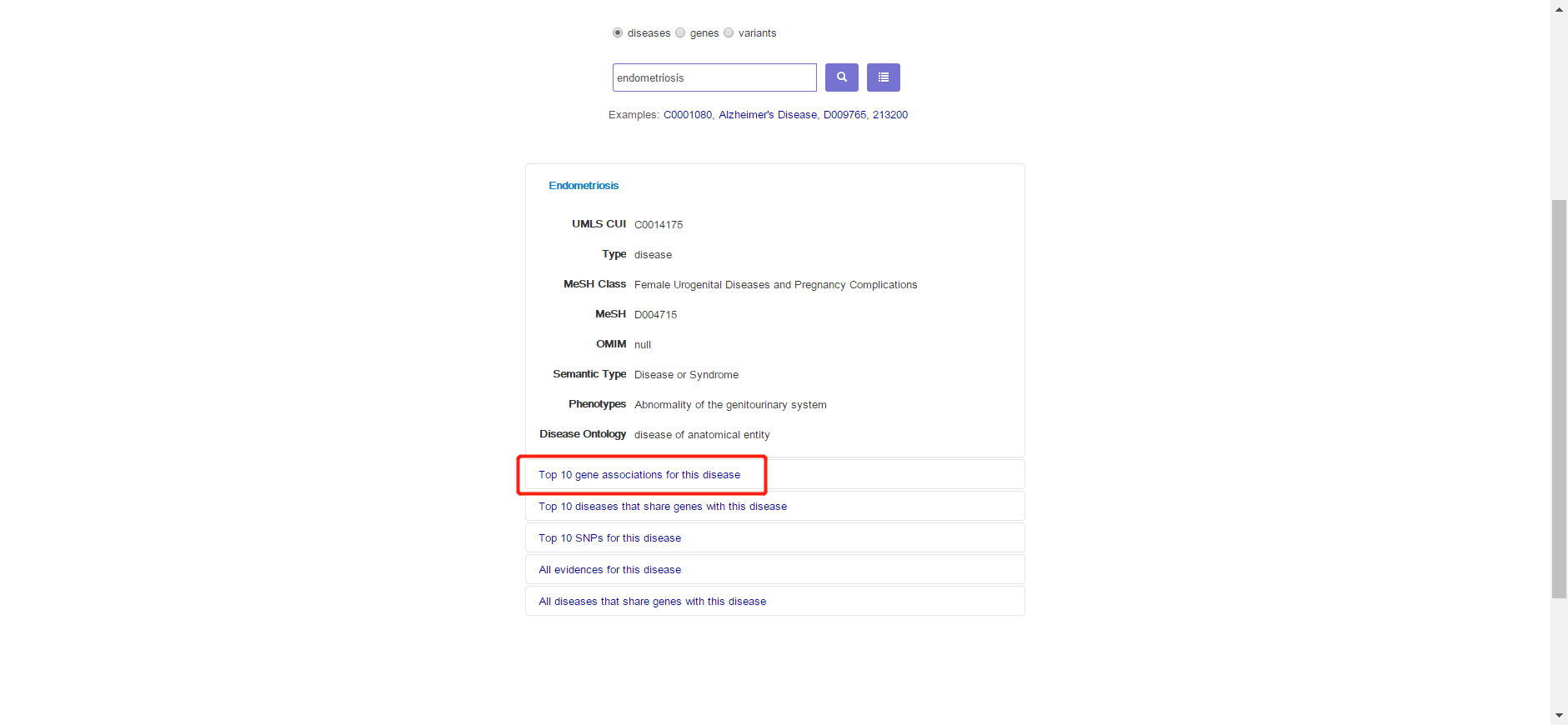
### DisGeNET

地址：<http://www.disgenet.org/web/DisGeNET/menu/search?1>

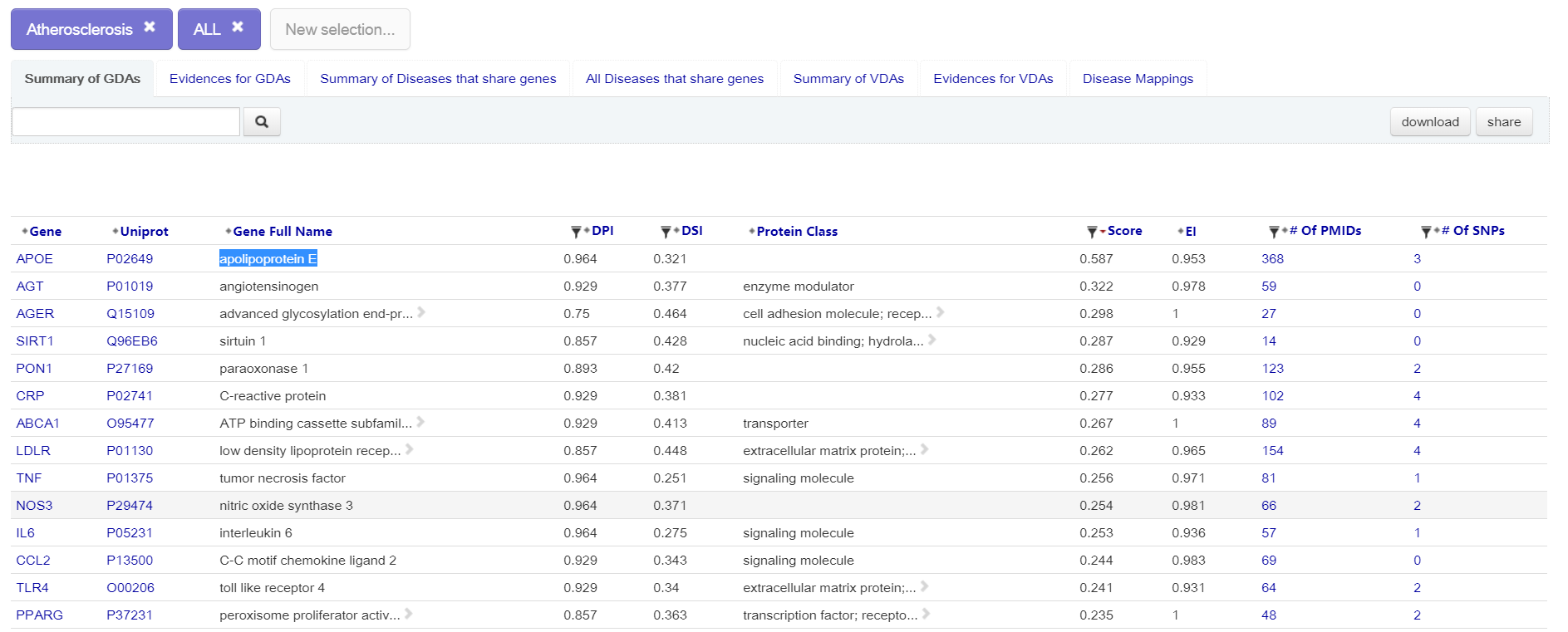
该数据库的使用是，在页面中输入疾病的名称，



输入疾病之后，查询搜索按钮，会得到对应的靶点数据，



这里存在很多的靶点信息，然后需要做的就是查询详细信息，下面是这些把靶点的信息。



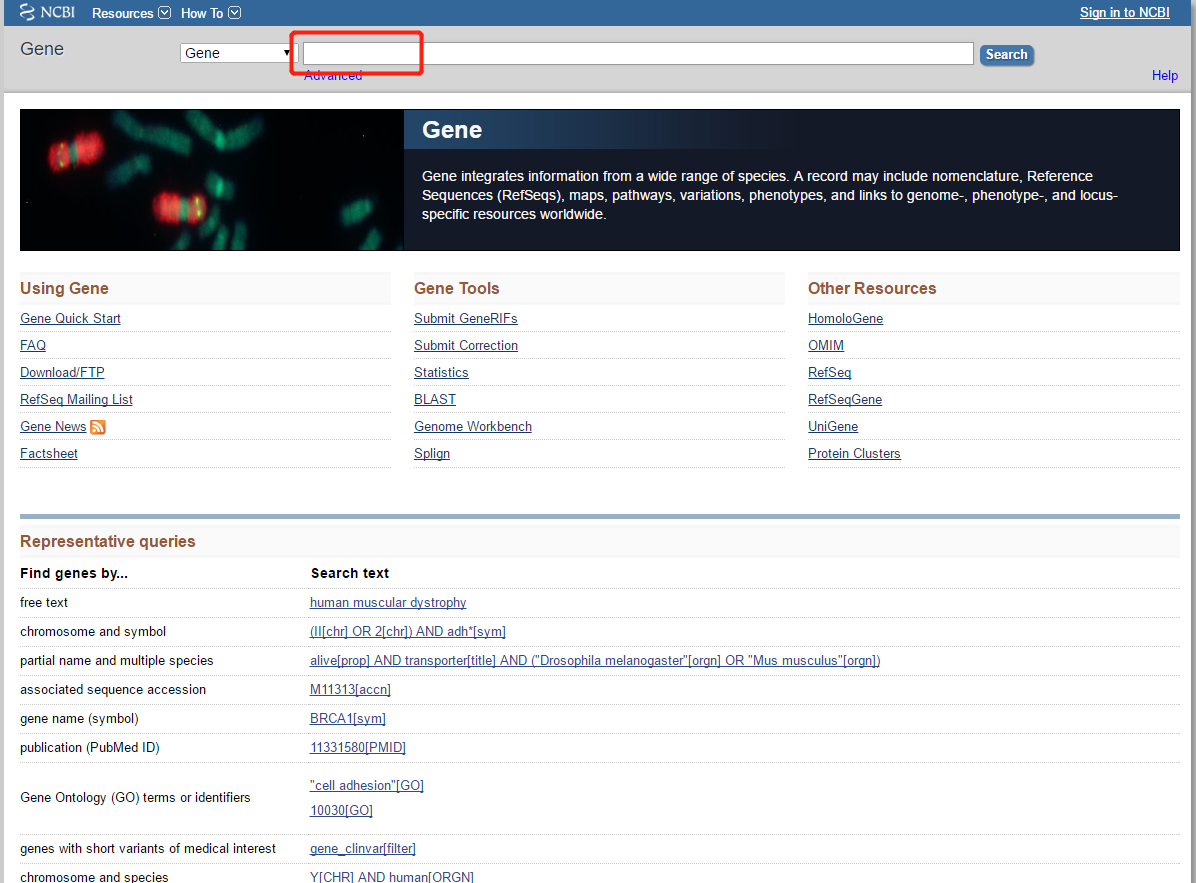
这里边的靶点进行筛选即可。

### 相关靶点获取

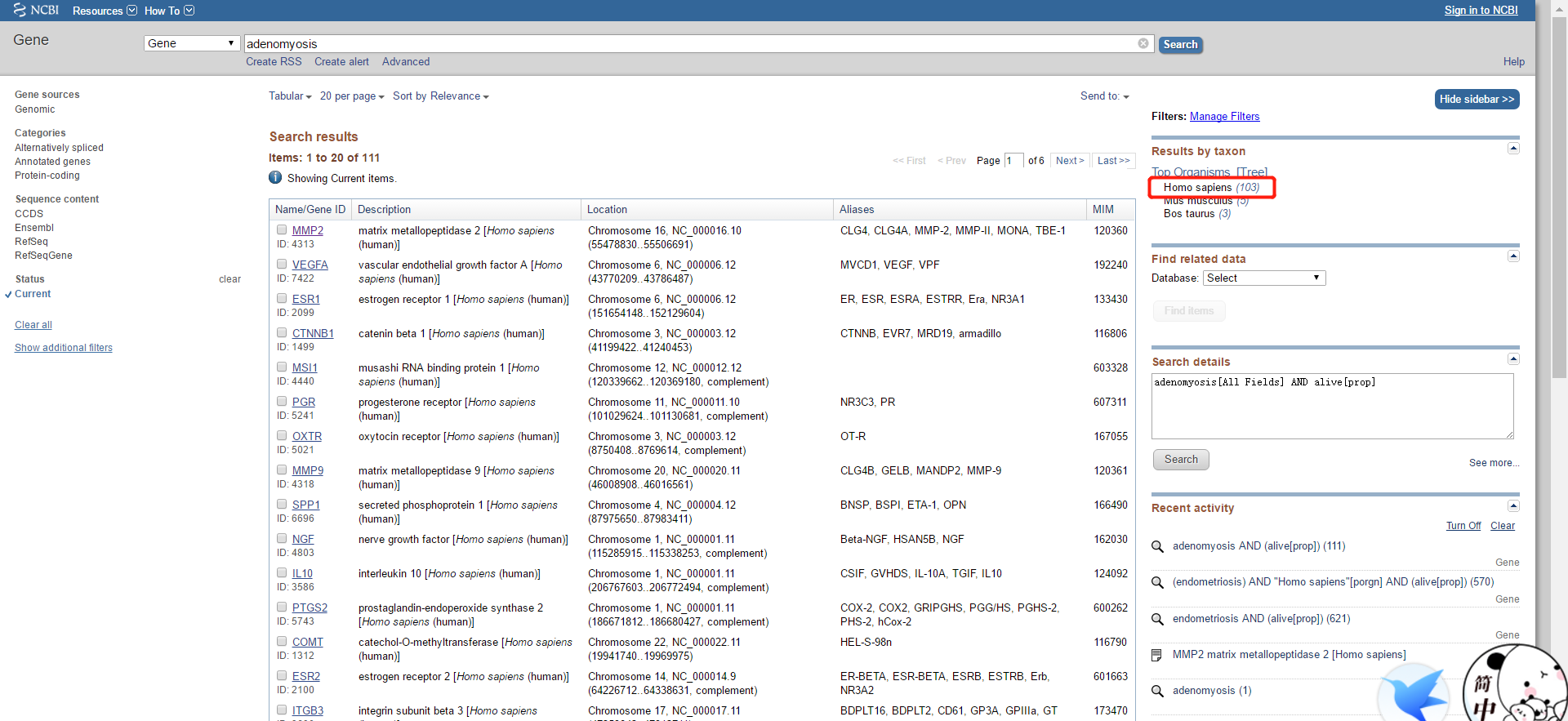
这里把靶点的获取分成两个部分，分别是使用DisGeNET这个数据库来获取疾病对应的靶点，这个很简单，但是另外有些成分是不能使用这个数据库获取到对应的靶点的，针对这种情况，现在需要做的就是使用NCBI获取到基因信息，在从Uniprot数据库里边数据到对应的靶点PDB ID.

NCBI获取基因信息。

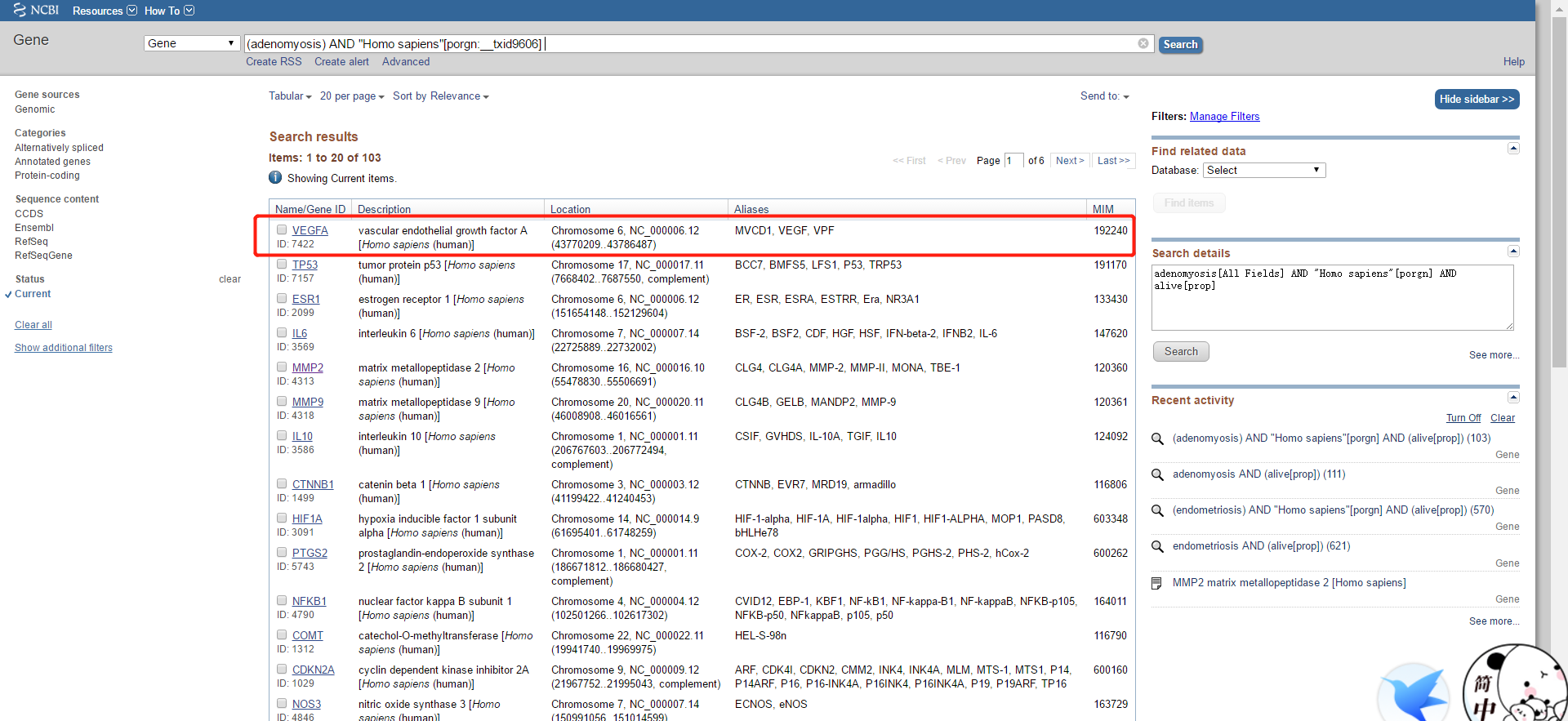
首先，在网址<https://www.ncbi.nlm.nih.gov/gene>对应的页面中，输入疾病信息，



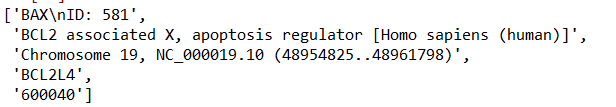
例如这里输入疾病adenomyosis，得到如下的结果，在结果中进行筛选，选择Home也就是



人类的基因，进行筛选后得到人类基因数据。

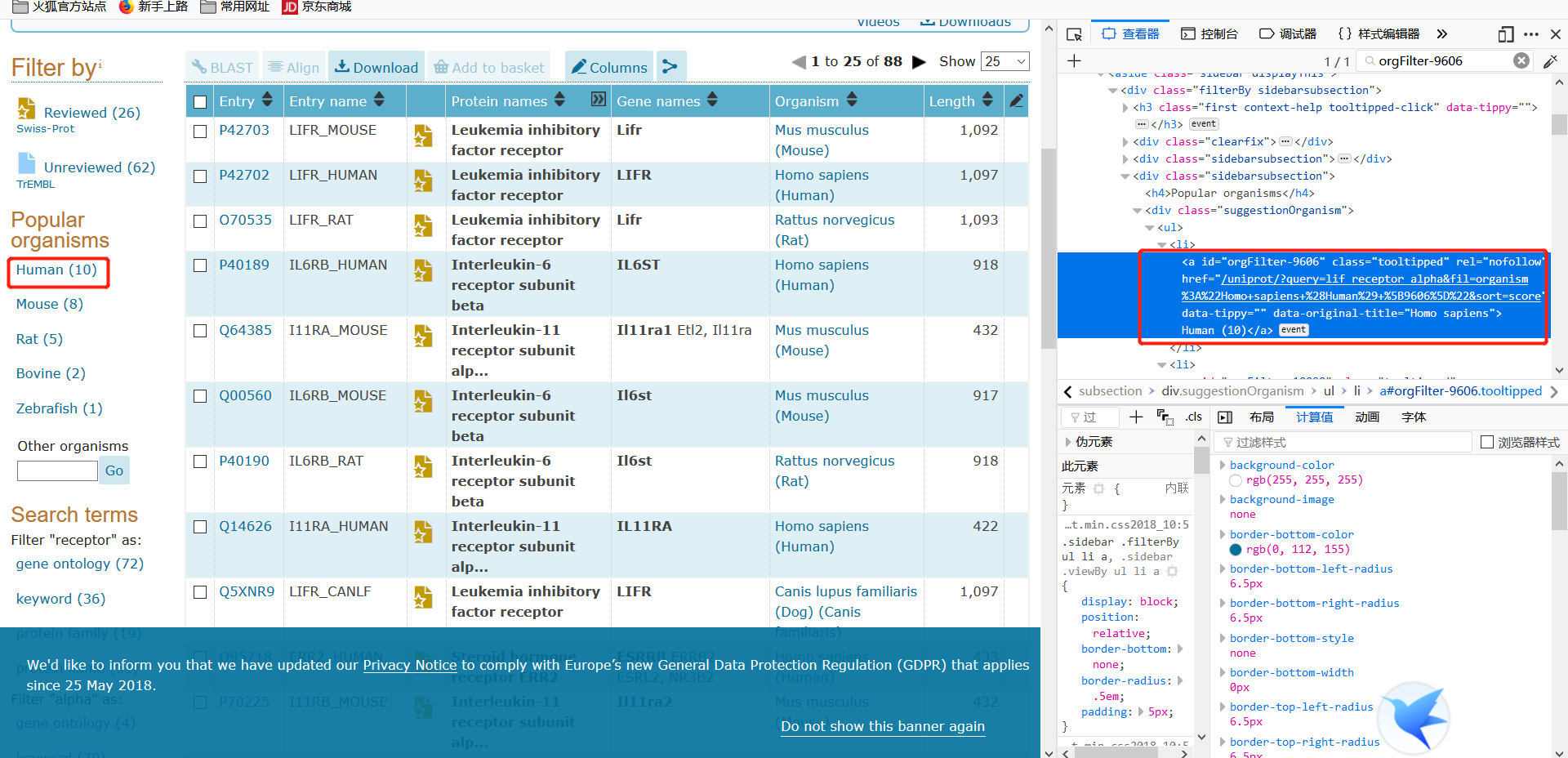


针对这样的筛选结果，下面其实要做的是把合适的结果存起来，现在看爬下来的数据，



查看到上面的数据，其实可以根据uniprot的数据可以知道，首先，本文需要获得的是基因靶点的ID,然后就是基因的名称，例如上面的BCL2和apoptosis regulator这两个数据是需要获取到的，处理的思路显然是对于后文的[Homo sapiens (human)]是务必需要去掉的，然后就是第一列的ID等信息务必除去，这样就能得到真正想要的数据。

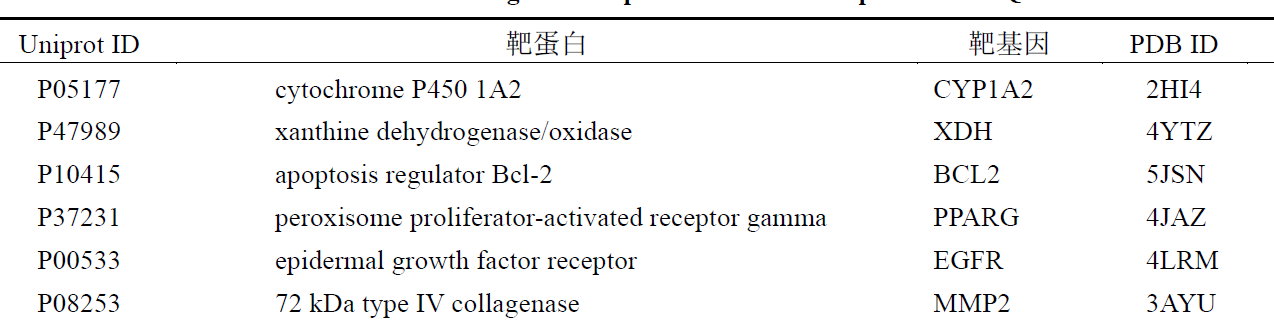
但是，现在发现，直接使用click操作会出现一些问题，导致点击操作实际上没用，针对这个种情况，发现上面的[Homo sapiens (human)]这部分，只要这个元素对应的序列中存在Human就可以认为这个元素是可以的，否则不是Human的就删去就好了。

另外一个就是Uniprot这个数据库，这个数据库在输入靶点的名称的时候，会出现下面的Human,但是右侧的分析可以发现，这个页面部分对应的href，然后我们只要获取这个href就能得到对应的查询页面，这个比上述的方法还要好一些。

存在异常的PDB

https://www.uniprot.org/uniprot/O43347

上述的数据获取到后，还需要格式化的数据，下面是格式化数据的



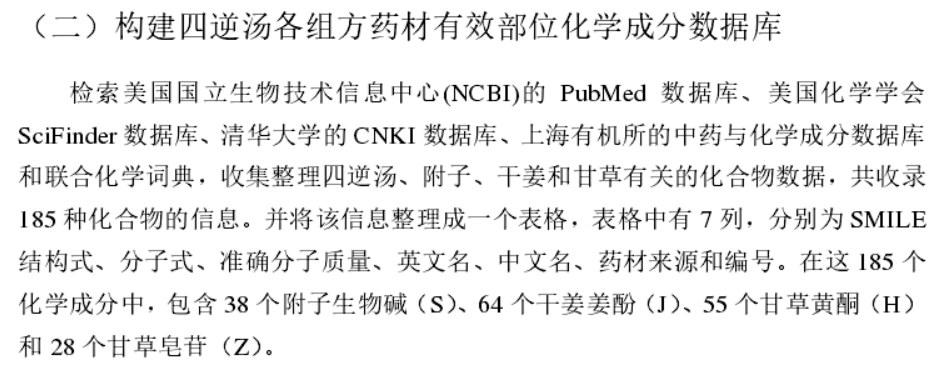
现在的情况是，数据必须是上面的格式才是可以的。

### 数据准备

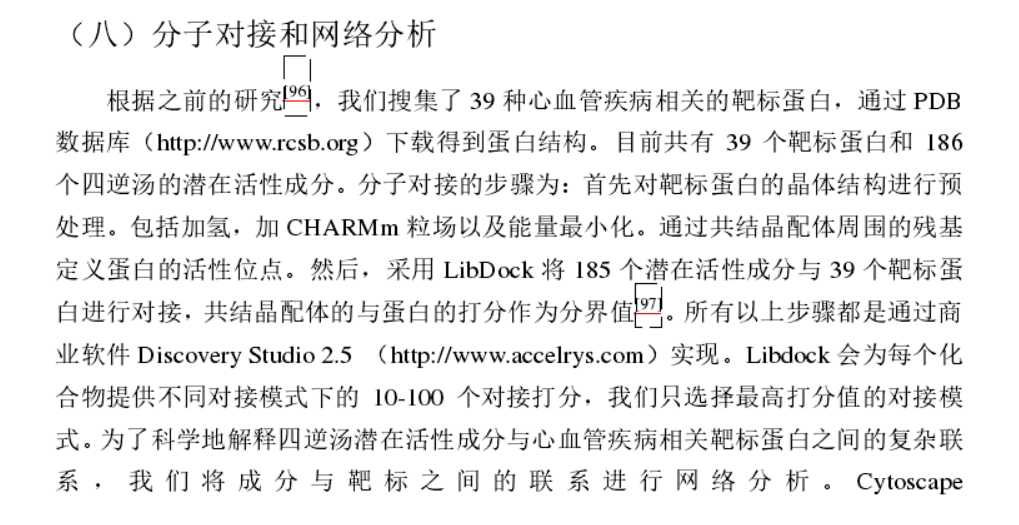
分子对接和虚拟筛选技术，需要准备小分子和靶点蛋白质，小分子数据的准备，某些参考文献中介绍了相关的方法[1],主要是从相关数据库里获取到相关的靶点数据。下面是参考文献中的一些介绍：

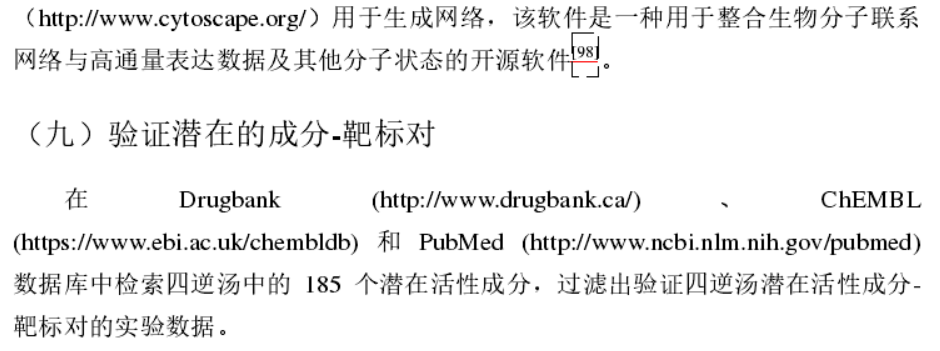
毒性与基因比较数据库CTD( Comparative Toxicogenomics Database，http://ctdbase. org/) 收集了化合物与基因、化合物与蛋白、化合物与疾病以及基因与疾病之间的相互作用关系。本研究使用CTD 数据库，以“Atherosclerosis”为关键词检索AS相关靶点，其中以“M”标记的为经过实验验证、可信度较高的疾病靶点。为保证实验结果的可靠性，本研究仅选择该类靶点，共计49 个。根据靶点的基因名称在PDB 网站( http: / /www. rcsb. org) 中查找其对应的蛋白质三维结构，并选择分辨率高的晶体结构，下载获得pdb文件共41个(表2)[2]

还有些文献的制备是如下所示的,这里介绍了化学成分数据库的准备。[3]

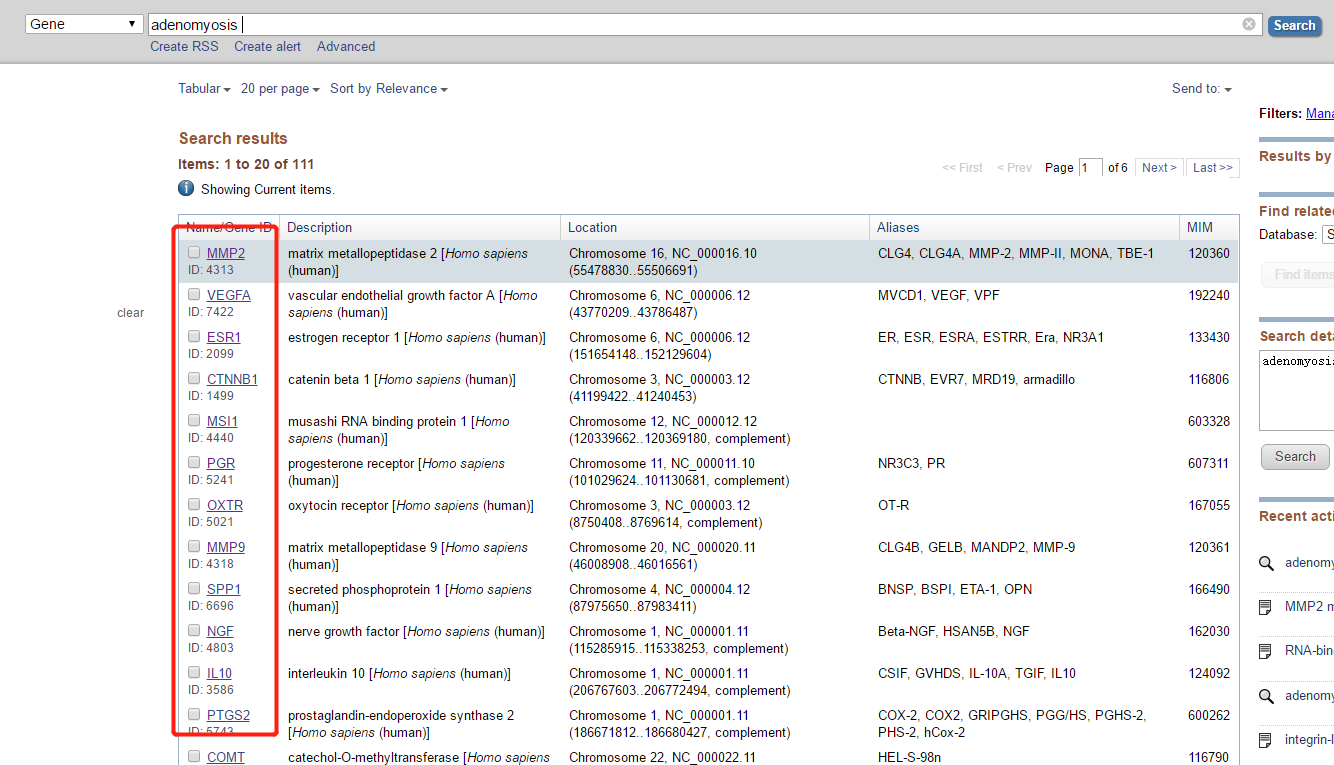


上述部分是成分的准备，下面的是介绍的网络对接和网络分析技术

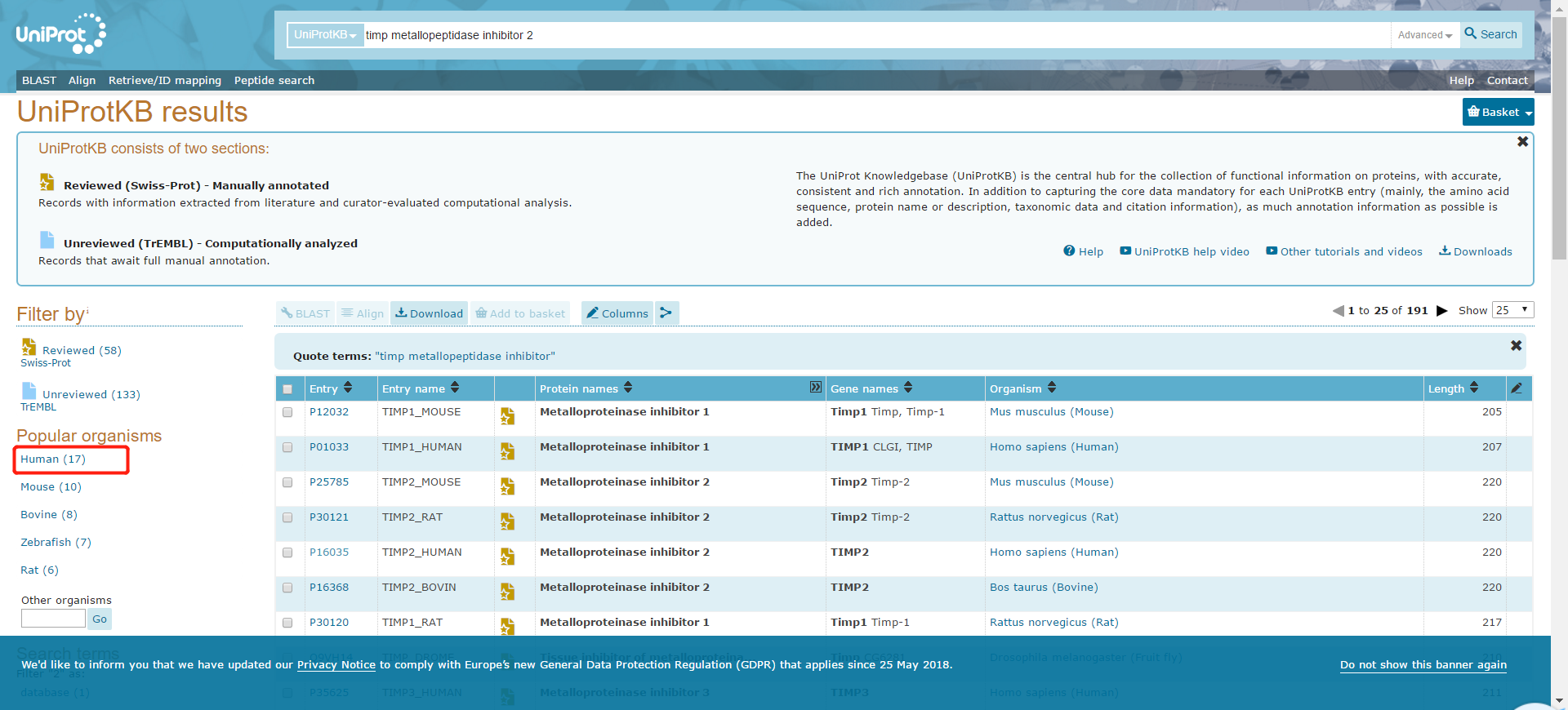




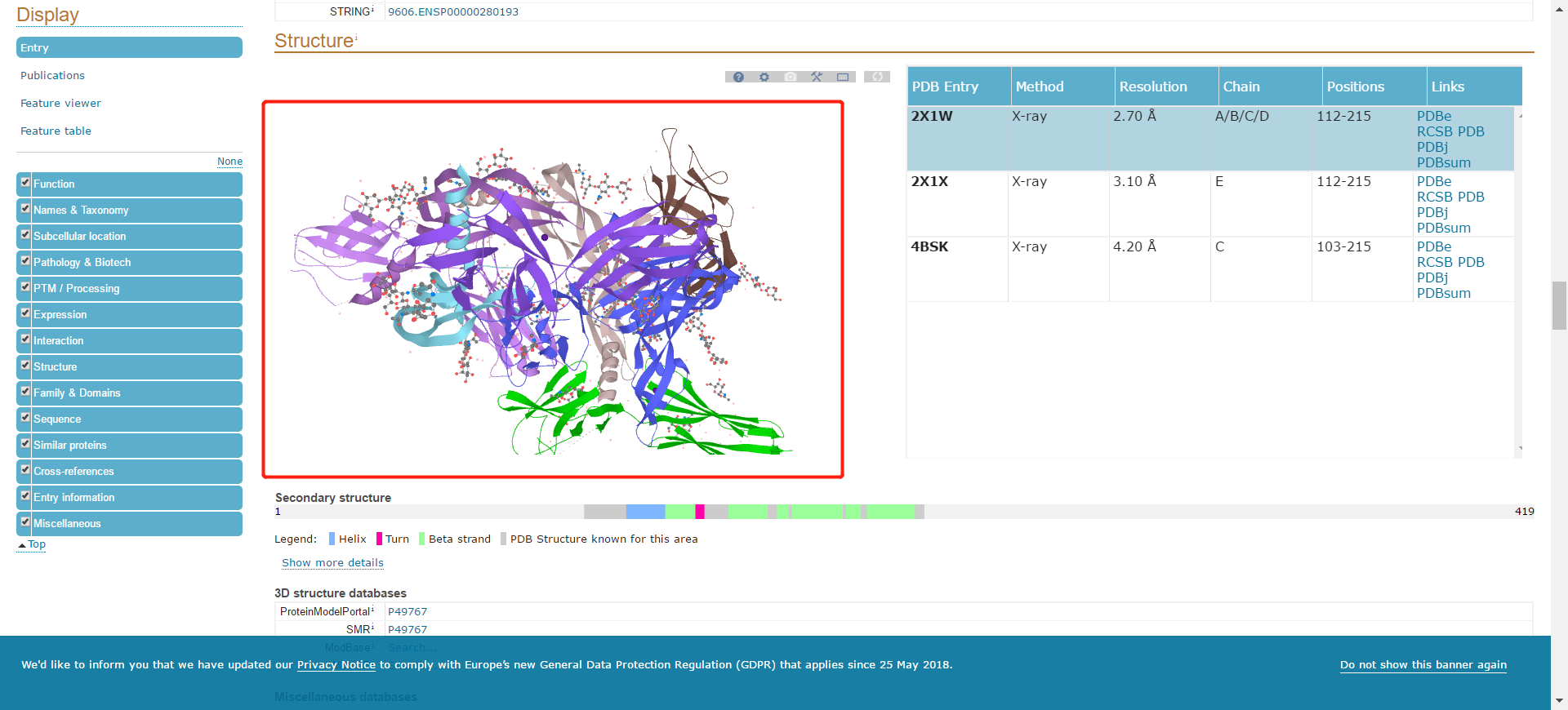
现在根据上述的信息，需要稍微做的就是，首选在NCBI中，输入关键词，然后搜索gene选项，找到基因信息，



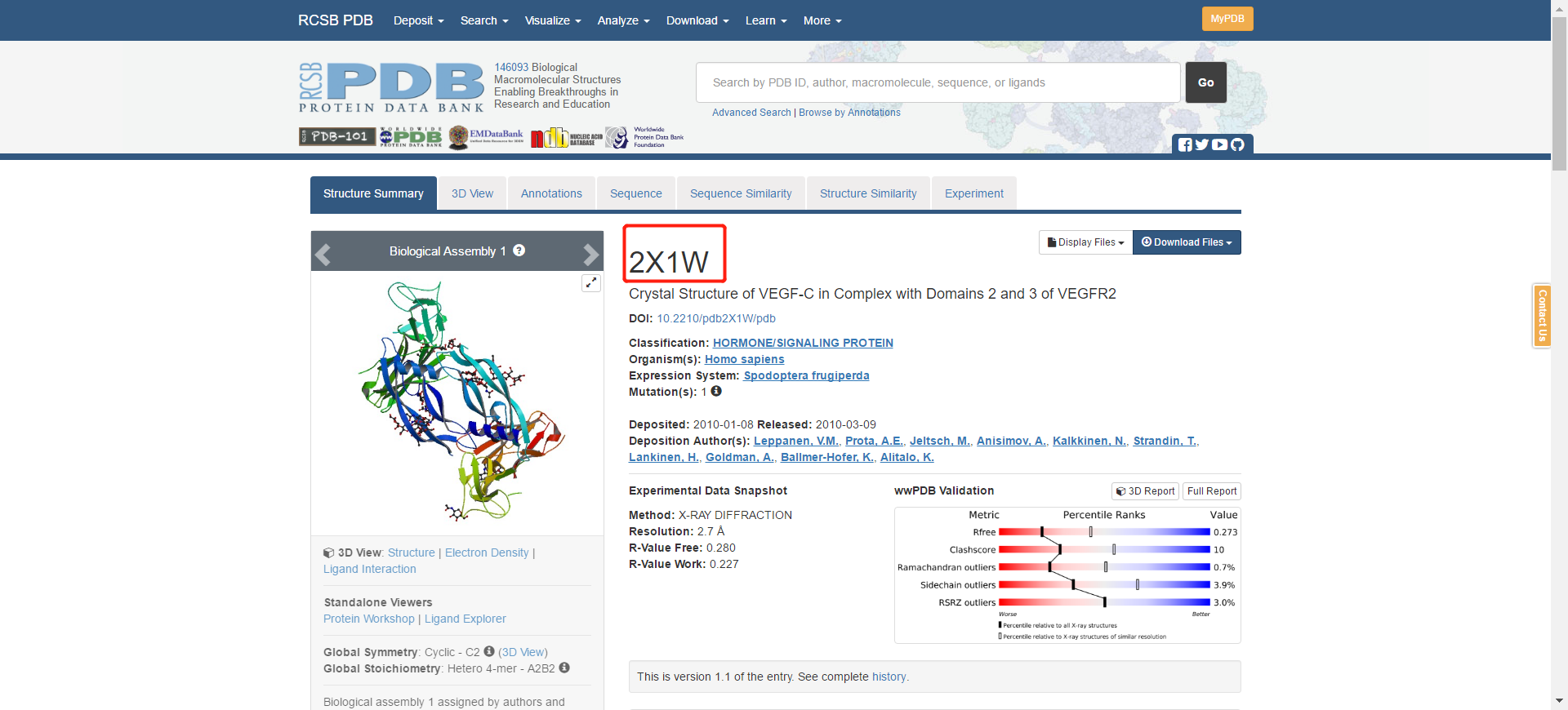
基因对应的有名称，然后根据这个名称到uniprot这个数据库里边根据基因的名称查找对应的靶点信息，



根据名称筛选HUMAN，然后找到对应的蛋白质，



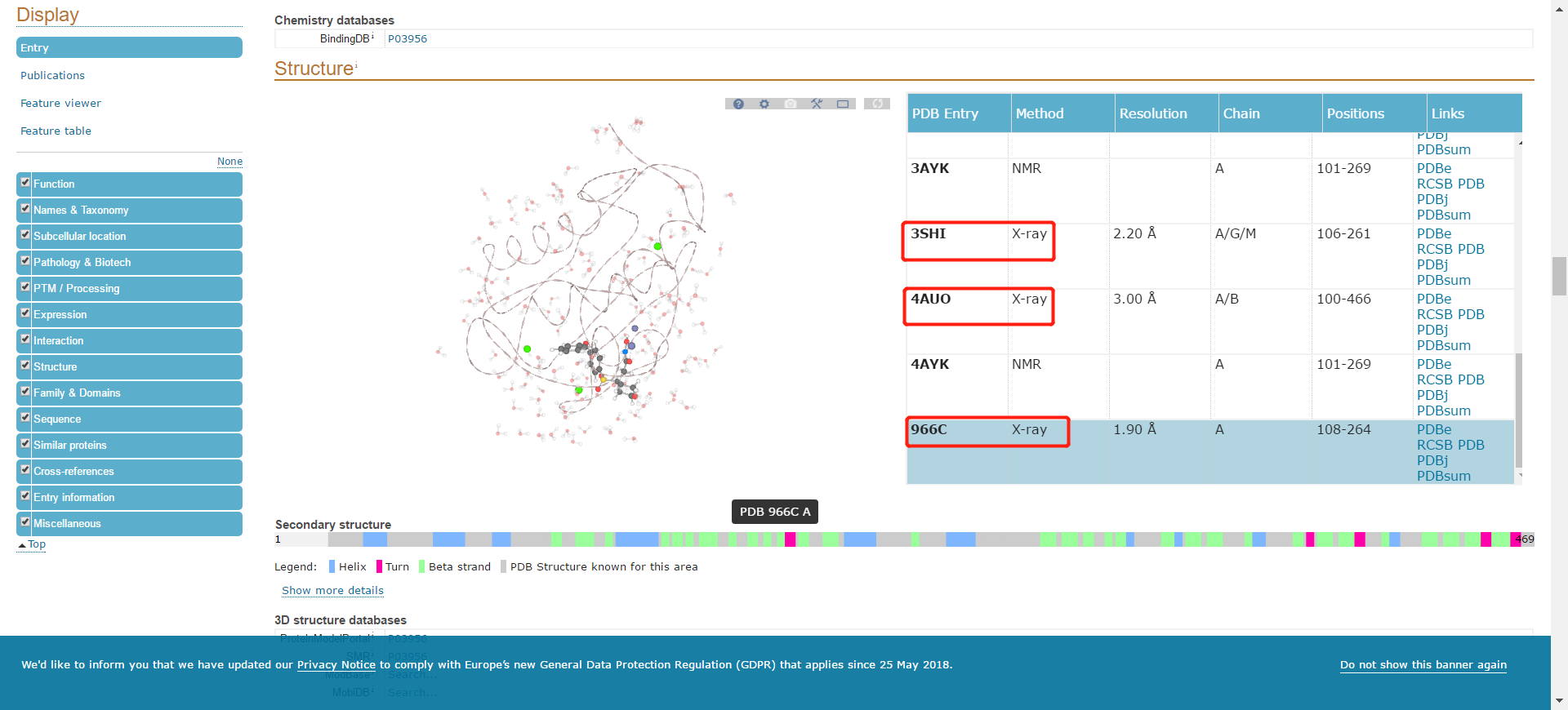
再将蛋白质的PDB对应ID找到，到PDB数据库里边找到对应的蛋白质信息。



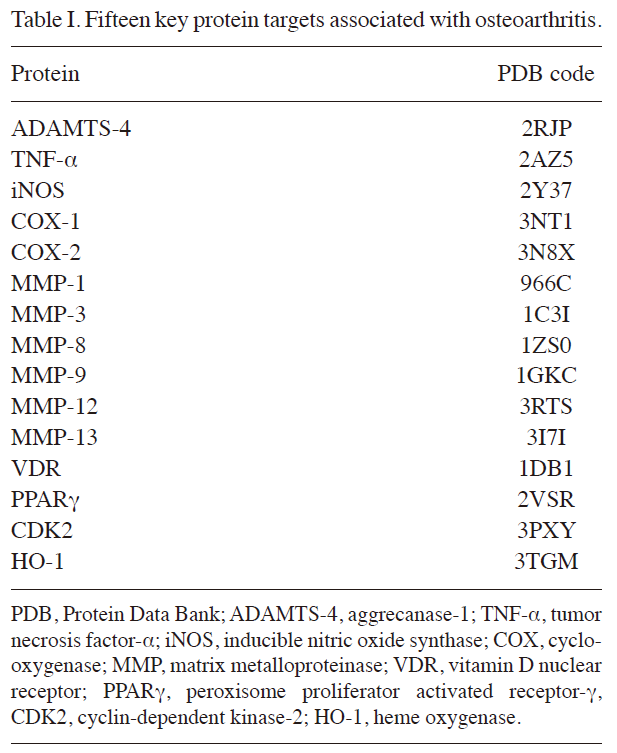
根据上述的分析，本文对靶点的数据进行进一步的分析

#### 存在的问题

现在在进行靶点筛选的时候，遇到几个问题。首先，给出了靶点的标号，例如，这里给出了一个靶点的编号MMP-1，试图在uniprot数据库里边查找这样一个对应的数据，在寻找的时候发现并不是所有的部分都是有结果的



但是，在相关文献中，这里直接给出的是966C,但实际上这里给出的不是这样的，下面是原始的给出的靶点蛋白质的名称和对应的PDB ID信息。



#### 靶点准备的思考

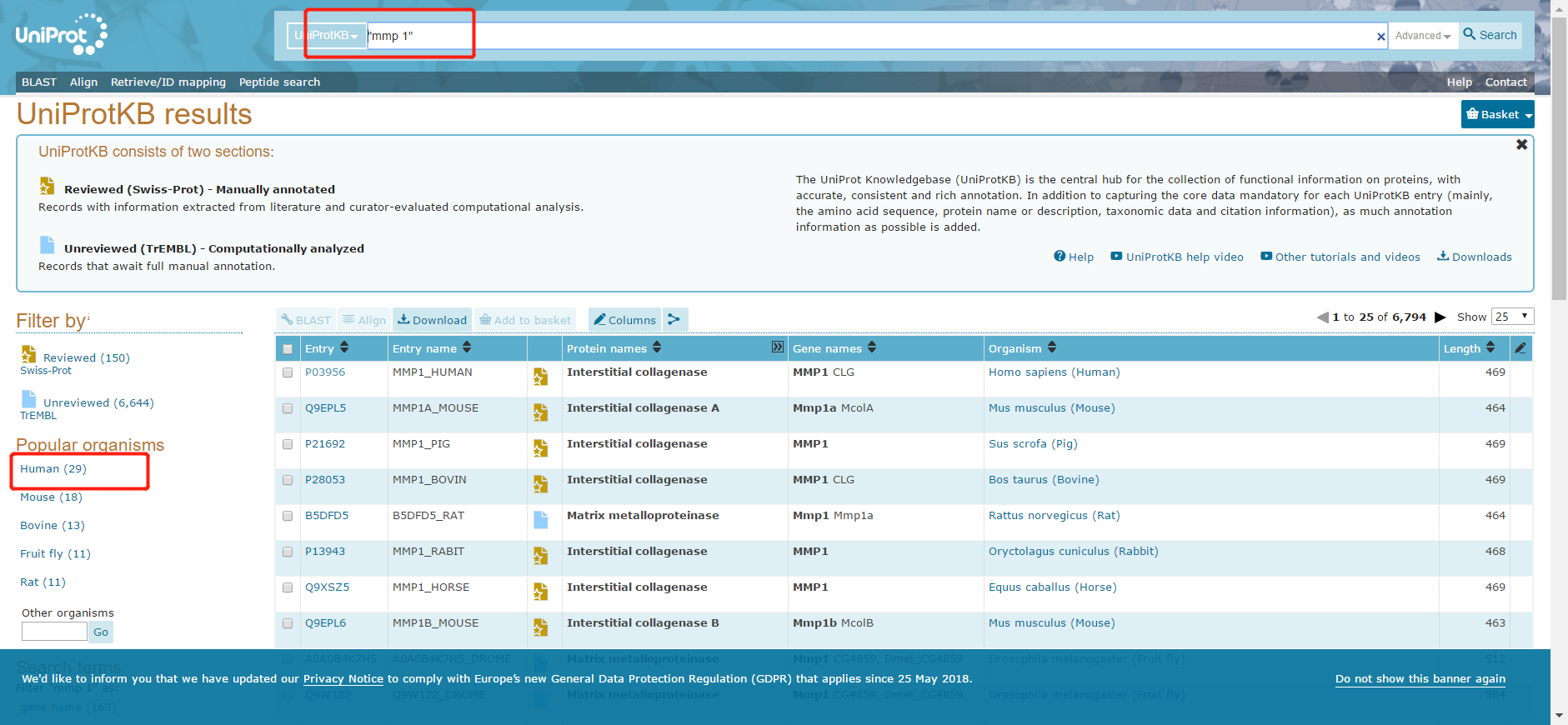
现在的研究发现，某种疾病对应的基因靶点是比较多的，前面提到的存在的问题中其实就介绍过一些。但是，除了上述的问题之外，有一点就是，这些靶点真正能用的是哪些呢，这是一个关键的问题，现在参考了一些[4]文献，这些文献介绍了一些常见的靶点，但是这里存在一个问题就是，靶点实际上是预先选择的。这个是需要注意的，另外一点需要注意的是，这里的靶点并非是每一个疾病对应很多的靶点，这种情况就很难处理。所以，现在的现状是靶点的选择是使用自动化的方式还是直接筛选的方式。

#### 靶点选择

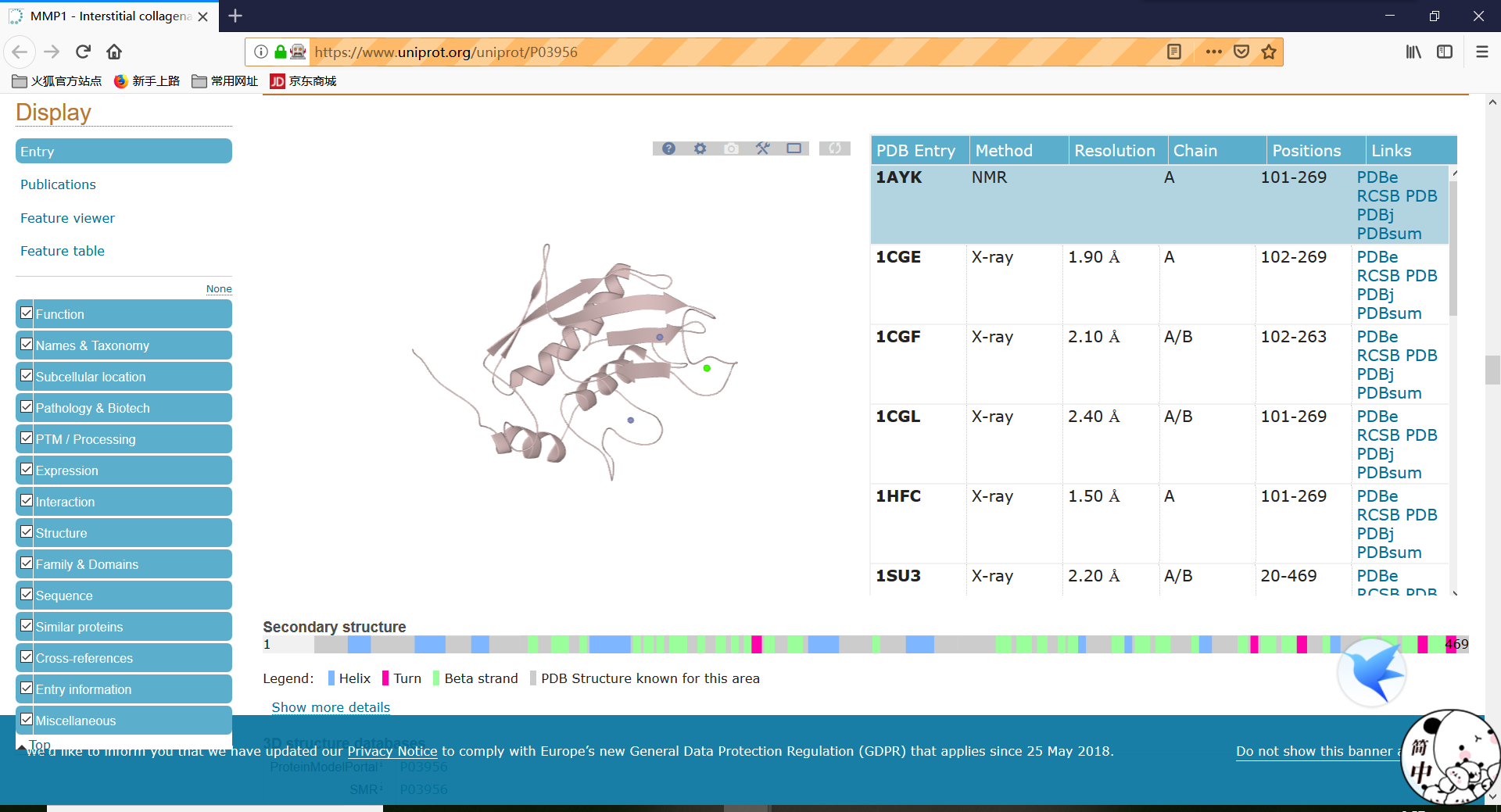
现在根据相关的资料，上述查询靶点中存在的问题通过查阅相关的文献了解到“同一个靶点对应的不同的PDB ID，实际上他们都是同一蛋白，只是在不同环境、不同分辨率、不同研究组或不同的其他条件下测得的结构”。这就是说，这里查询到的不同蛋白质是其实是同一种蛋白质，只是需要选择对应的PDB ID才好，现在的情况是，给定了靶点的名字，找到对应的靶点的数据。下面，对这个算法进行设计。下面是靶点蛋白的汇总。



总体的逻辑是这样的。首先，需要根据查询到的靶点的名称，输入到UniProt数据库中，地址是<https://www.uniprot.org/>，例如靶点MMP-1,如下所示:



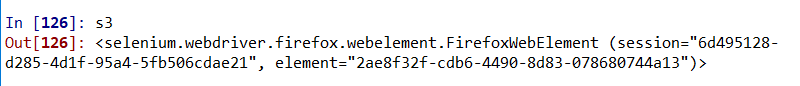
接下来，选择Human，然后再选择对应的Entry ,选择后点进去会进入到下面的页面，这个



页面中存在的很多的蛋白质信息，但是实际上需要选择的是Method为X-ray,Resolution按照从小到大的排序，取较小的一个。这里需要的这三个数据，稍微做一下排序就好。

根据上面的算法，下面是涉及的





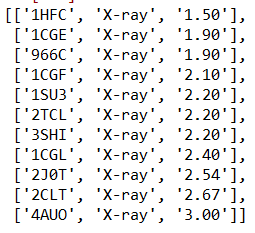
另外，需要注意的是这里有三个页面的代码需要处理，分别是当前的，前一个页面的代码，现在发现由于存在iframe框架，也就是说那个页面实际上是页面里边内嵌入的页面，这就导致不能直接进入到小页面中进行确定位置，所以处理的方式就是使用selenium的页面转换就是进入到子页面进行选择。最终定位到对应的页面，获取到页面中的PDB数据。然后需要对这些数据进筛选，筛选的算法是：

输入的数据：一个二维数组，存储着PDB的相关数据

1. 仅仅要前三行数据
2. 选择Method为X-ray的数据
3. 去掉Resolution的单位
4. 按照Resolution进行排序

输出的数据：排好序的一个二维数组，按照一定的格式进行排序的结果。

使用相关的案例结果如下所示：



接下来的一个处理的思路是查询页面的设计，查询页面的设计是类似的设计思路，主要是输入靶点的名称获取到Human对应的PDB的查询页面的地址。

## 对接过程分析

前面的数据处理之后，现在有两个疾病，分别是adenomyosis，endometriosis，对这两个疾病对应的靶点取交集，除去那些没有对应的靶点蛋白的数据，现在总共可以获取到59个基因靶点，分别如下所示

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 3AYU | 3FQN | 2HYV | 3U9Q | 4I8V | 2HAZ |
| 1MKK | 1SQN | 1BR9 | 1XW6 | 4OEE | 3V8S |
| 3CBP | 6ESM | 5JFW | 4DRI | 1HY7 | 5NRR |
| 3FQN | 3CXD | 4H6J | 2W96 | 3V96 | 4D0Z |
| 3FQN | 5F19 | 1SVC | 2W96 | 5IRQ | 1RG8 |
| 1SQN | 3BWY | 1BI7 | 1DOK | 1IKN | 3VH8 |
| 3AYU | 3OLL | 4D1P | 2QFA | 2Y6D | 1B6U |
| 1MKK | 1MIZ | 2Z62 | 5L0R | 5HHE | 2J0I |
| 3CBP | 1HFC | 4O1V | 3KYO | 3FXZ | 1B8K |
| 3FQN | 4ZIE | 4XDX | 2R7G | 4QBS |  |

接下来对上面的靶点进行对接操作，另外，这里根据成分数据库，首先整理得到两个数据库，分别是TCMSP和TCMID数据库得到的成分数据。

## 对接结果分析

在进行对接的时候，出现了一个结果，结果可以看到，如下的情形，下面是类似于教科书式的条形图显示，但是，另外一个就是前面的Result Table的数据，这里把Result Table的结果给爬虫到本地，最好是处理到Mysql数据库中，然后从数据库中使用sql脚本进行筛选。

Sql的设计是这样的，首先就是查看数据的情况

|  |
| --- |
|  |
| 对接结果数据呈现 |

上面的结果实际上给了进行数据库设计的典型思路，设计id,Proteins,PDBID,testCompounds,

Score这几个方面的表名为result的表，这里的id是自动增长的。另外，还可以设计维表，

|  |
| --- |
|  |
| 对接结果可视化 |

主要是为了存储id和成分的名称数据，这个可以给后边的可视化进行帮助，除此之外，还有一个需要做的就是这些数据结合前面的进行药-成分-靶点可视化操作。

表设计result

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键ID | int(11) | 11 | PRI |
| Proteins | 蛋白质 | varchar(32) | 20 |  |
| Pdb | PDB名称 | varchar(32) | 20 |  |
| Compound | 化合物 | varchar(32) | 20 |  |
| Score | 打分 | varchar (32) |  |  |

建表语句：

CREATE TABLE result(id int(11) primary key auto\_increment,Proteins varchar(32),Pdb varchar(32),Compound varchar(32),Score varchar(32));

除了上面的数据之外，还要设计维表，主要是成分数据和对应的化合物名称的关联关系，下面是对应的成分表设计medicine,然后后边是建表语句。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段 | 中文名称 | 数据类型 | 长度 | 索引 |
| id | 主键id | int(11) | 11 | PRI |
| compound | 化合物编号 | varchar(32) | 20 |  |
| name | 化合物名称 | varchar(256) | 256 |  |
| drug | 药品名称 | varchar(100) | 100 |  |

建表语句：

CREATE TABLE medicine(id int(11) primary key auto\_increment,compound varchar(32),name varchar(256),drug varchar(100));

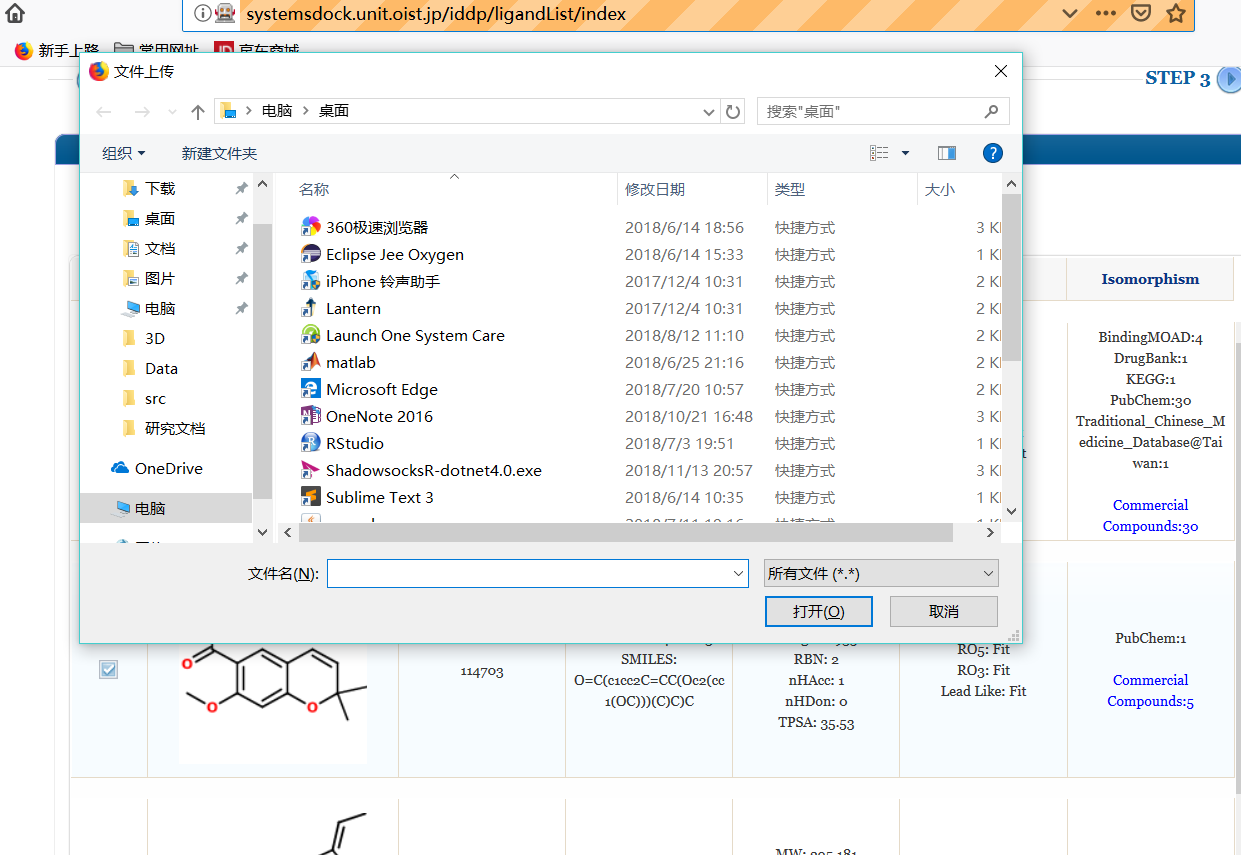
上面的两张表可以进行药品-成分-靶点的分析。另外由于表result中存储着对应的打分，这就表明，这里可以根据表里的打分进行进行筛选，选出打分是合理的结果进行再使用Cytoscape软件进行数据的分析。

除了上述的对接设计外，后续还有一些东西，主要是进行后续页面数据的爬取,现在的情况是如何进行这个页面的定位。但是，我们发现这个数据实际上是可以下载到本地的，页面中存在这样的下载按钮，

## 对接中的问题

在对接操作中遇到一些问题，这些问题也是后期需要解决的方面。

1.对接中需要上传文件，但是这里我们发现上传文件遇到下面的页面不能关闭的问题，下面是对接中遇到的页面，这个页面的情况就是，页面出现一个“文件上传”小框，但是这个框的话在整个的操作完成后是不能进行关闭的，搜索发现针对这种情况，是需要使用另外的一种工具来做的，那就是autoit，这个小工具后期可以进行修改，完善脚本的使用。



2.在进行自动化操作的时候，中间会遇到一些问题，那就是需要一些异常处理机制，在进行开发的时候，异常处理机制是十分需要的，这里有一个问题就是，在进行页面访问的时候需要判断这个是不是可以直接找到JS执行的时长，如果能够找到执行的时长，那么就能够不需要给定执行的时长，而是直接根据JS执行的时长确定下一步的操作。

3.前端页面需要进行一番的修改。在进行开发的时候，本实验发现，前端是一个需要注意的东西，现在的情况是，进行对接开发的时候，发现其实没有前端页面是比较烦的事情，这个时候可以考虑使用前端页面进行封装一下，使得程序的封装是比较完美的。这样的话就能比较好的显示出效果。

现在考虑的前端框架是JS或者是django框架进行相关的

## 富集相关分析

在完成对接的操作后，下面要做的就是进行基因富集的分析，这里进行富集操作使用的数据库是DAVID、STRING、MsigDB这两个数据库可以进行基因富集的分析。下面对这些分析的工程进行相关的介绍与分析。这里需要对富集分析进行一番的解释。所谓富集分析就是

### DAVID

### MsigDB

### STRING

参考文献:

[1]

[2] 基于分子对接及生物网络功能模块识别的复方丹参滴丸的网络药理学研究

[3]基于化学信息学的抗心衰相关药物筛选和作用机制探究讨论，陈思。

[4] 江振洲, 刘晓昕, 周旺,等. 我国疾病靶点研究最新进展[J]. 药学进展, 2015(5):335-350.