自动化网络药理论文构思

## 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。

无法自动化的部分：分子对接这个地方是不能够进行自动化的，因为这个部分的数据处理需要使用到特定的软件discovery studio来进行，是无法进行进一步的，原因在于这一步使用discovery studio软件的时候，需要在linux环境下进行安装和使用，但是这种情况是根本不可能进行的。所以根据这个情况本文的重点在于前期的数据准备上。前期的数据准备也是能出文章的关键点。

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

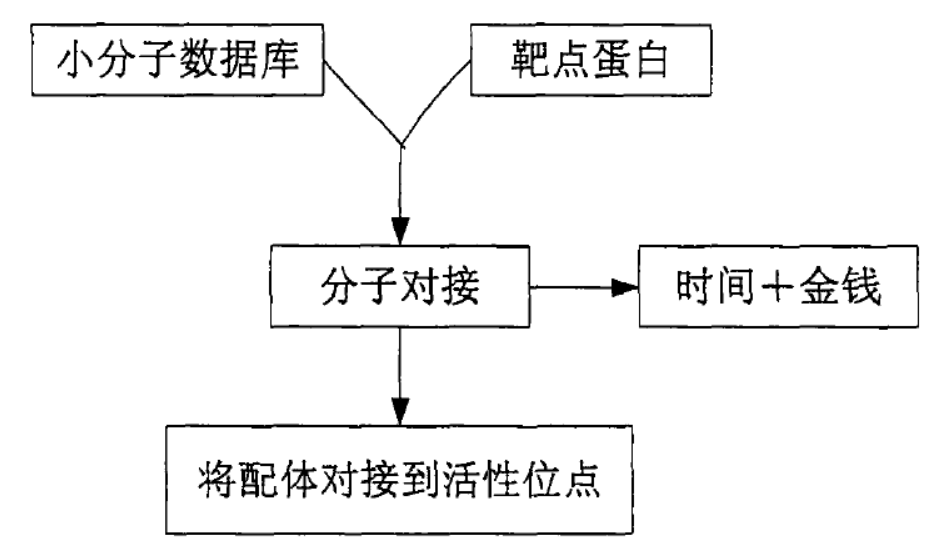
现在选择的数据库主要是两个数据库，TCMSP和TCMID这两个数据库，现在分别就这两个数据库进行相关的算法设计。软件设计的主要目的是为了准备小分子配体的结构信息而进行的相关的操作。整体思路是，使用selenium+python的技术组合搭建自动化环境，进行药物分子数据的获取。在具体的技术实现上，需要使用网页分析技术，分析具体页面的特点，使用selenium模拟人的点击操作，最终获取到对应的数据，使用python的DataFrame数据进行数据的存储，并最终将数据到处到。

### 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。

### 反向对接

近年来,一些科学家提出了反向分子对接(inverse docking)叫l的概念,反向分子对接是利用小分子去筛选一系列靶标蛋白,其目的也不再是筛药,而更多的是去探索药物作用机理或毒副作用机制。反向对接方法对于某些临床药物有较好的预测结果。[注:本文主要要做的不是对新药进行预测，而是解析药物的作用机制。针对这种情况，本文可以把反向对接的自动筛选算法作为一个关键来做]

### 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。

### 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。

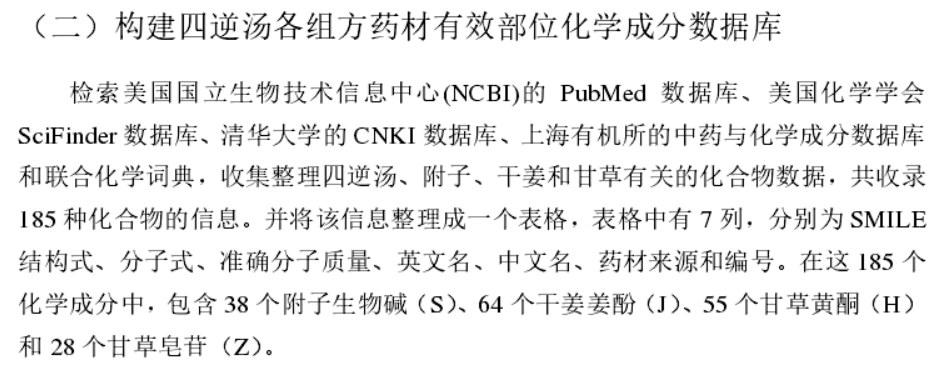
### 相关靶点获取

### 数据准备

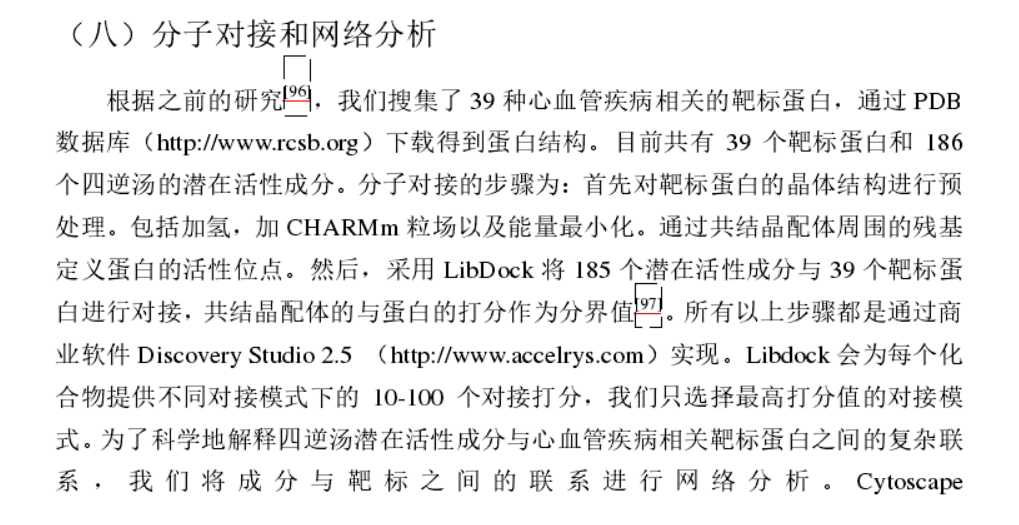
分子对接和虚拟筛选技术，需要准备小分子和靶点蛋白质，小分子数据的准备，某些参考文献中介绍了相关的方法[1],主要是从相关数据库里获取到相关的靶点数据。下面是参考文献中的一些介绍：

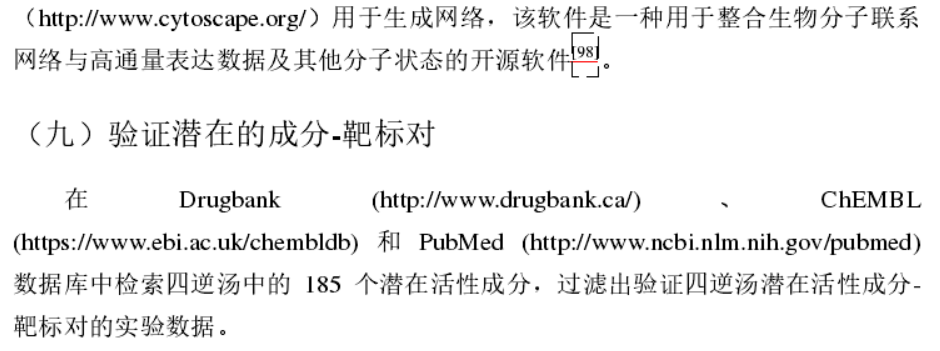
毒性与基因比较数据库CTD( Comparative Toxicogenomics Database，http://ctdbase. org/) 收集了化合物与基因、化合物与蛋白、化合物与疾病以及基因与疾病之间的相互作用关系。本研究使用CTD 数据库，以“Atherosclerosis”为关键词检索AS相关靶点，其中以“M”标记的为经过实验验证、可信度较高的疾病靶点。为保证实验结果的可靠性，本研究仅选择该类靶点，共计49 个。根据靶点的基因名称在PDB 网站( http: / /www. rcsb. org) 中查找其对应的蛋白质三维结构，并选择分辨率高的晶体结构，下载获得pdb文件共41个(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

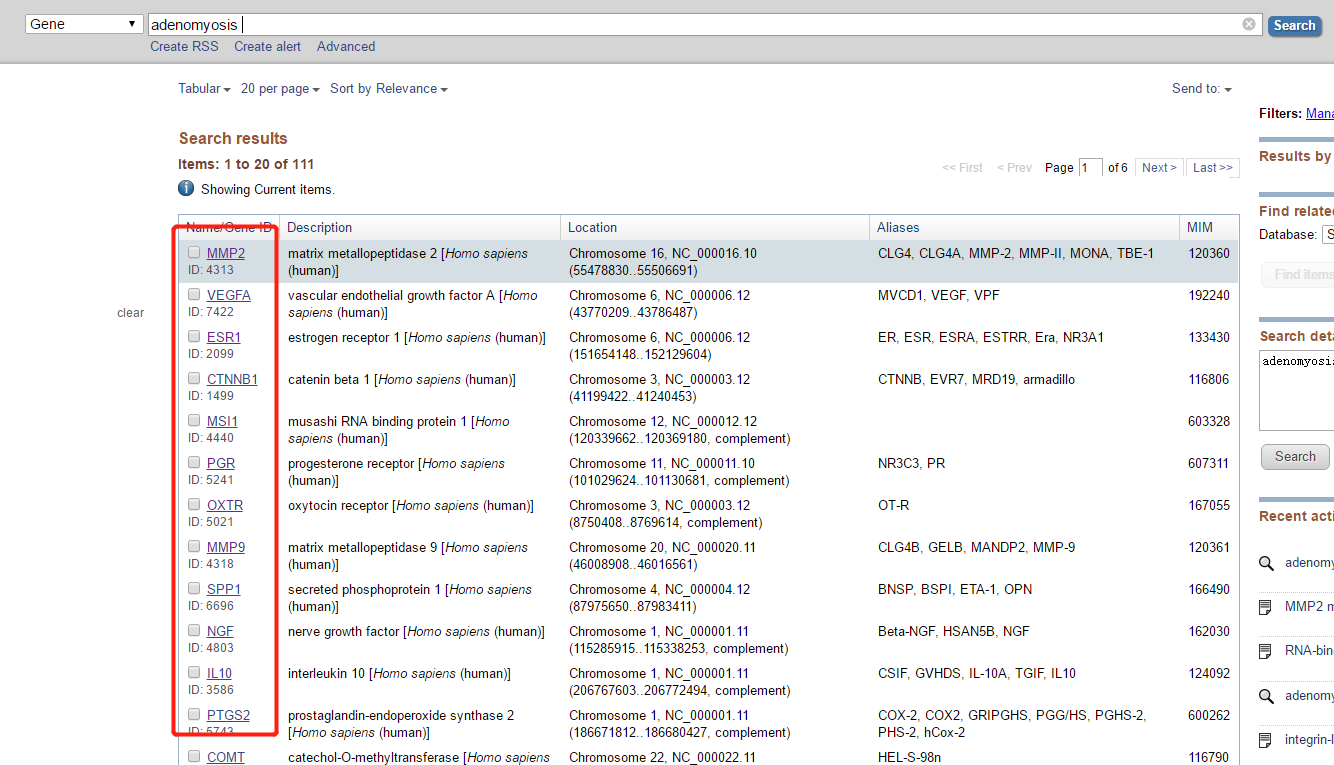


上述部分是成分的准备，下面的是介绍的网络对接和网络分析技术

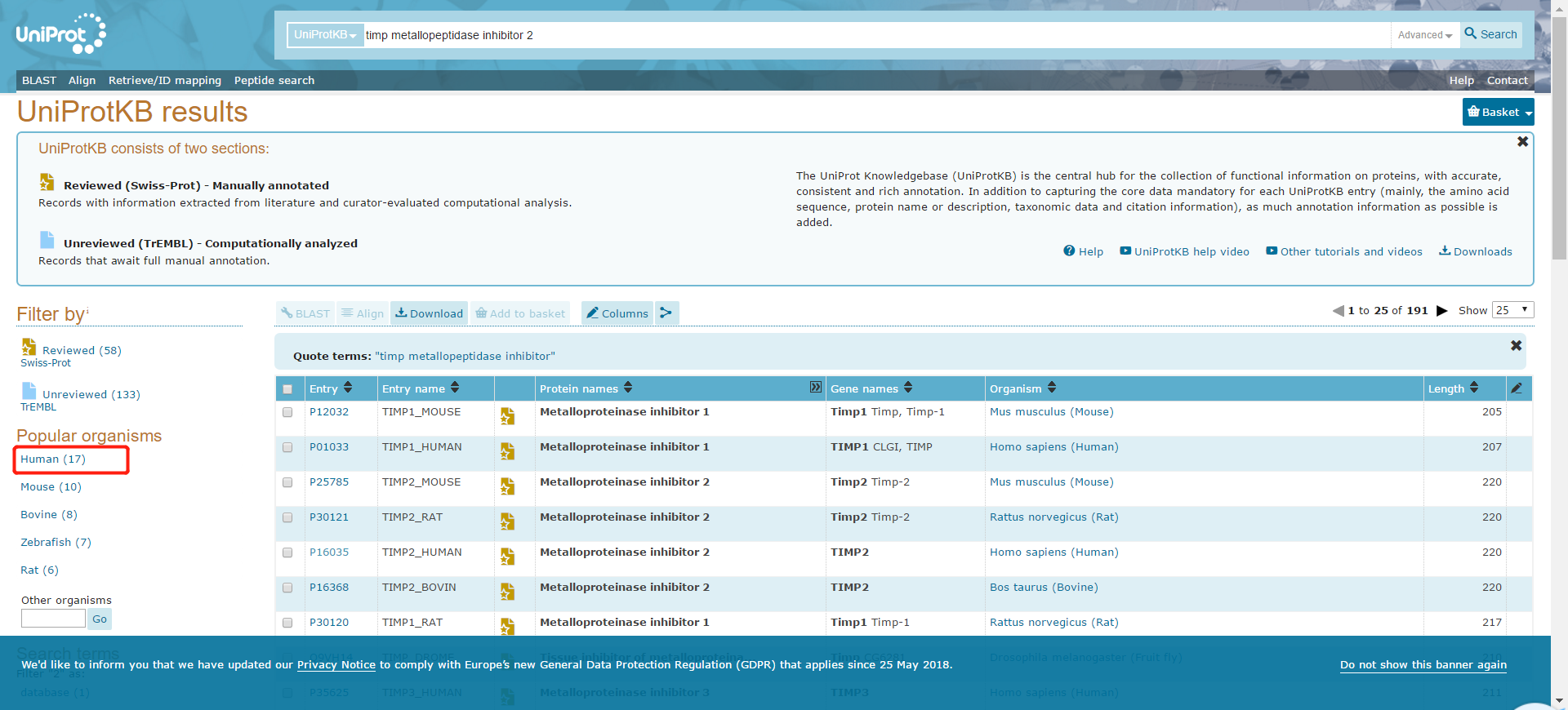




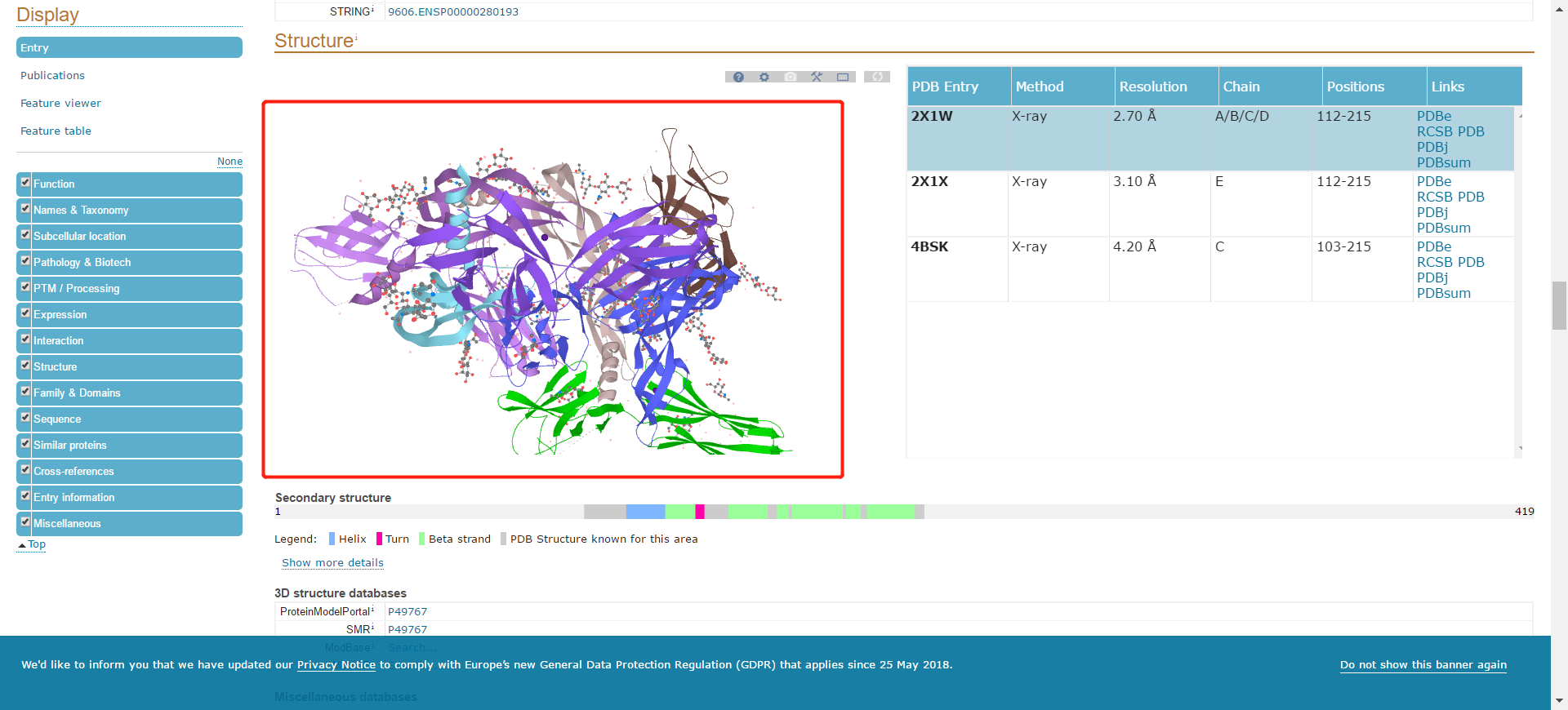
现在根据上述的信息，需要稍微做的就是，首选在NCBI中，输入关键词，然后搜索gene选项，找到基因信息，



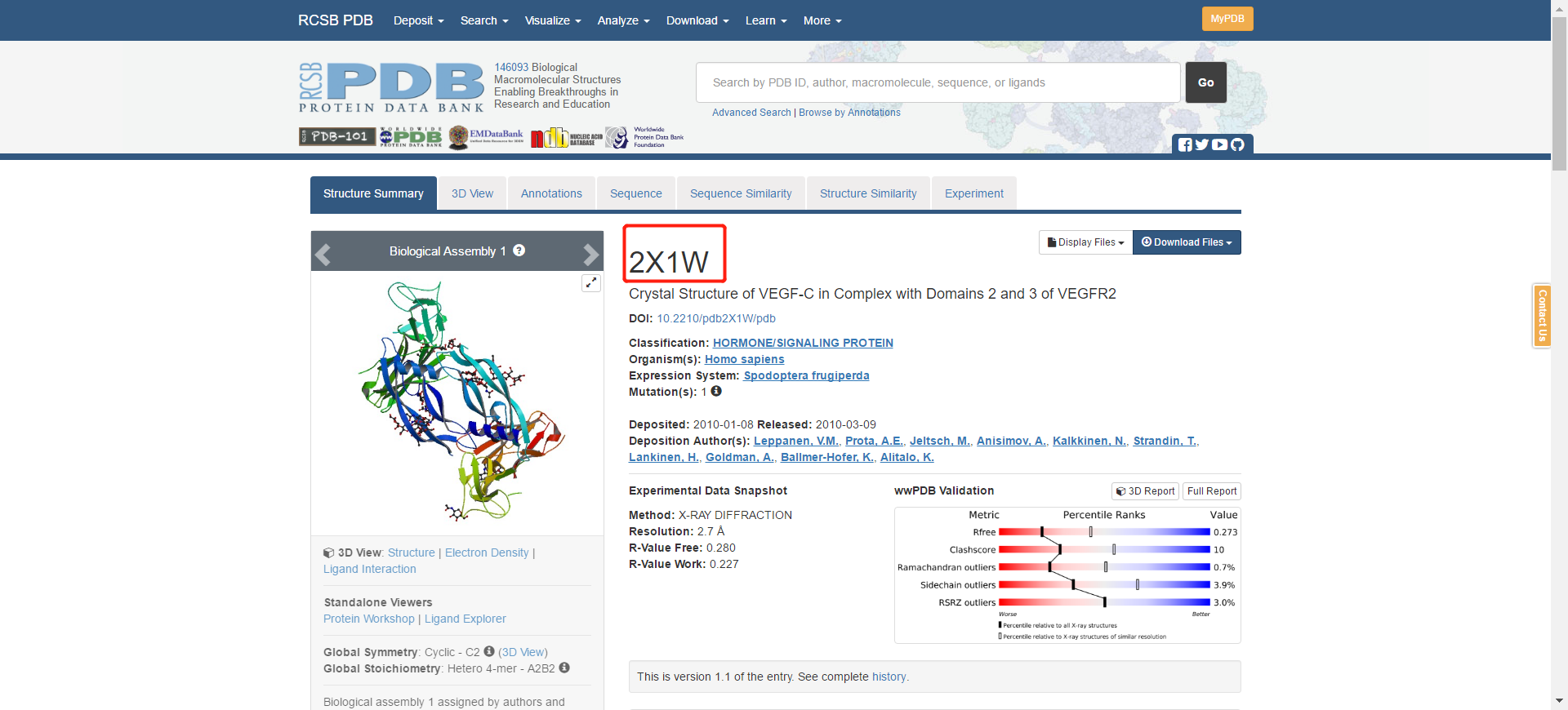
基因对应的有名称，然后根据这个名称到uniprot这个数据库里边根据基因的名称查找对应的靶点信息，



根据名称筛选HUMAN，然后找到对应的蛋白质，



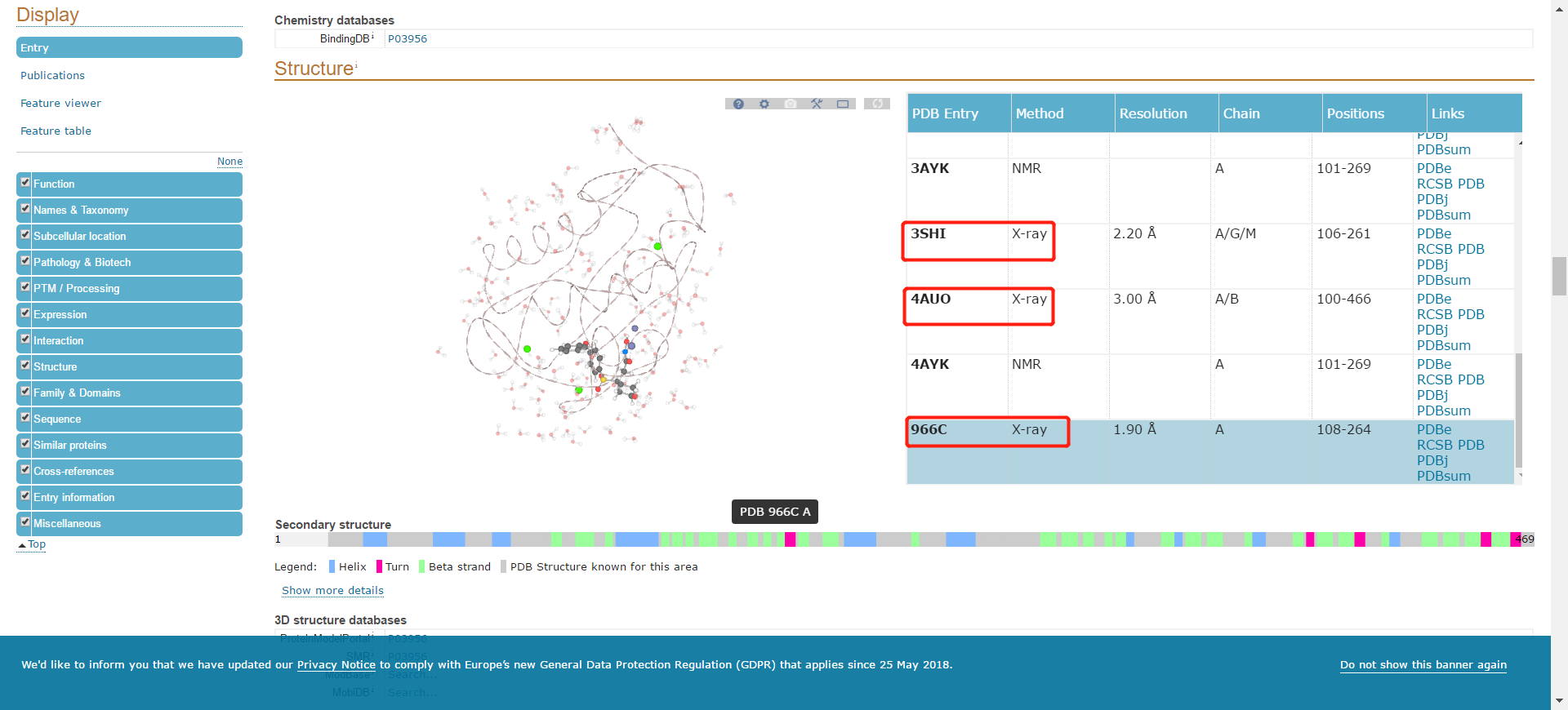
再将蛋白质的PDB对应ID找到，到PDB数据库里边找到对应的蛋白质信息。



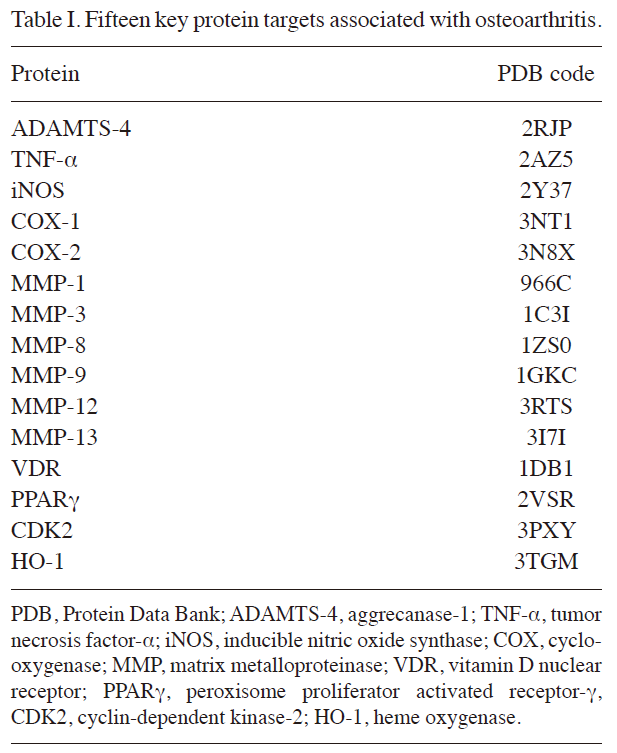
根据上述的分析，本文对靶点的数据进行进一步的分析

#### 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



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

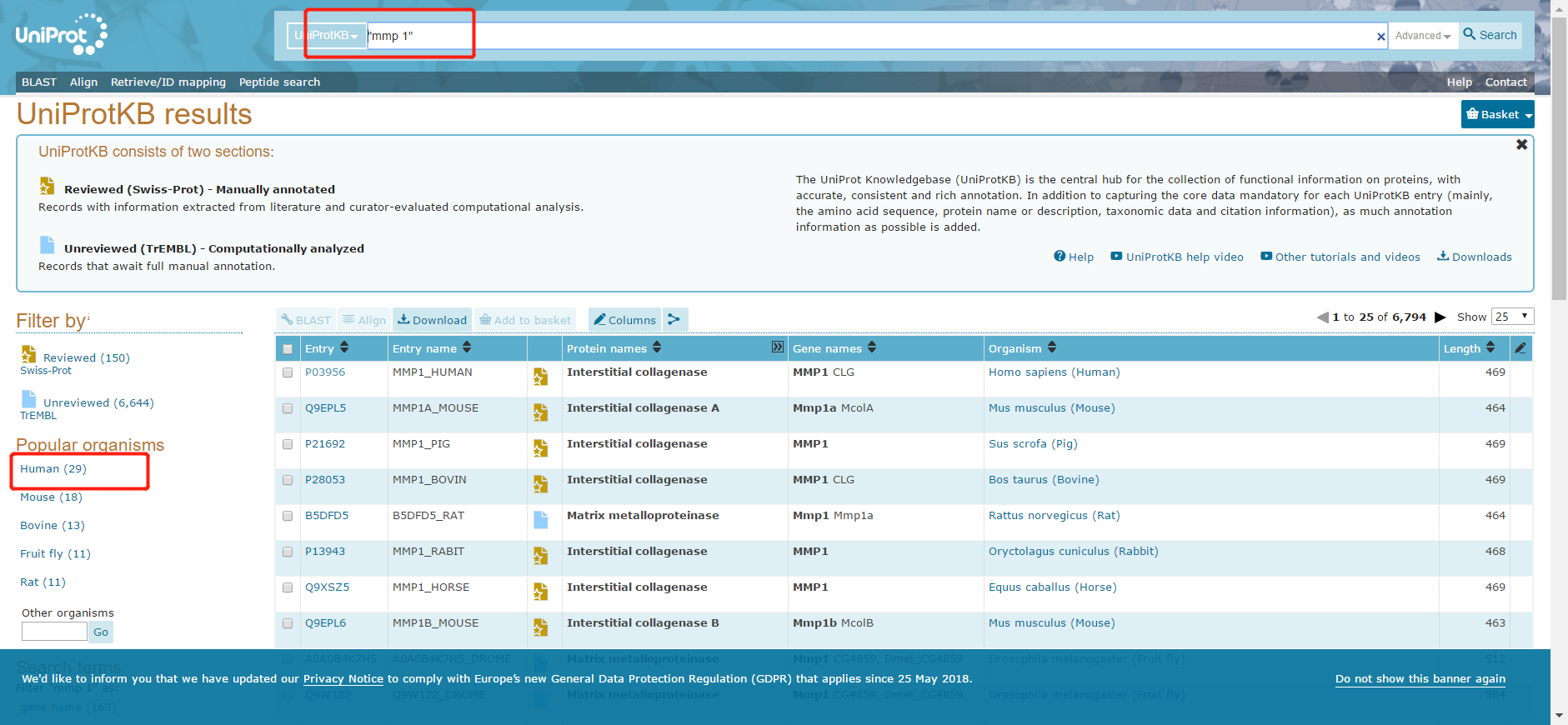
现在的研究发现，某种疾病对应的基因靶点是比较多的，前面提到的存在的问题中其实就介绍过一些。但是，除了上述的问题之外，有一点就是，这些靶点真正能用的是哪些呢，这是一个关键的问题，现在参考了一些[4]文献，这些文献介绍了一些常见的靶点，但是这里存在一个问题就是，靶点实际上是预先选择的。这个是需要注意的，另外一点需要注意的是，这里的靶点并非是每一个疾病对应很多的靶点，这种情况就很难处理。所以，现在的现状是靶点的选择是使用自动化的方式还是直接筛选的方式。

#### 靶点选择

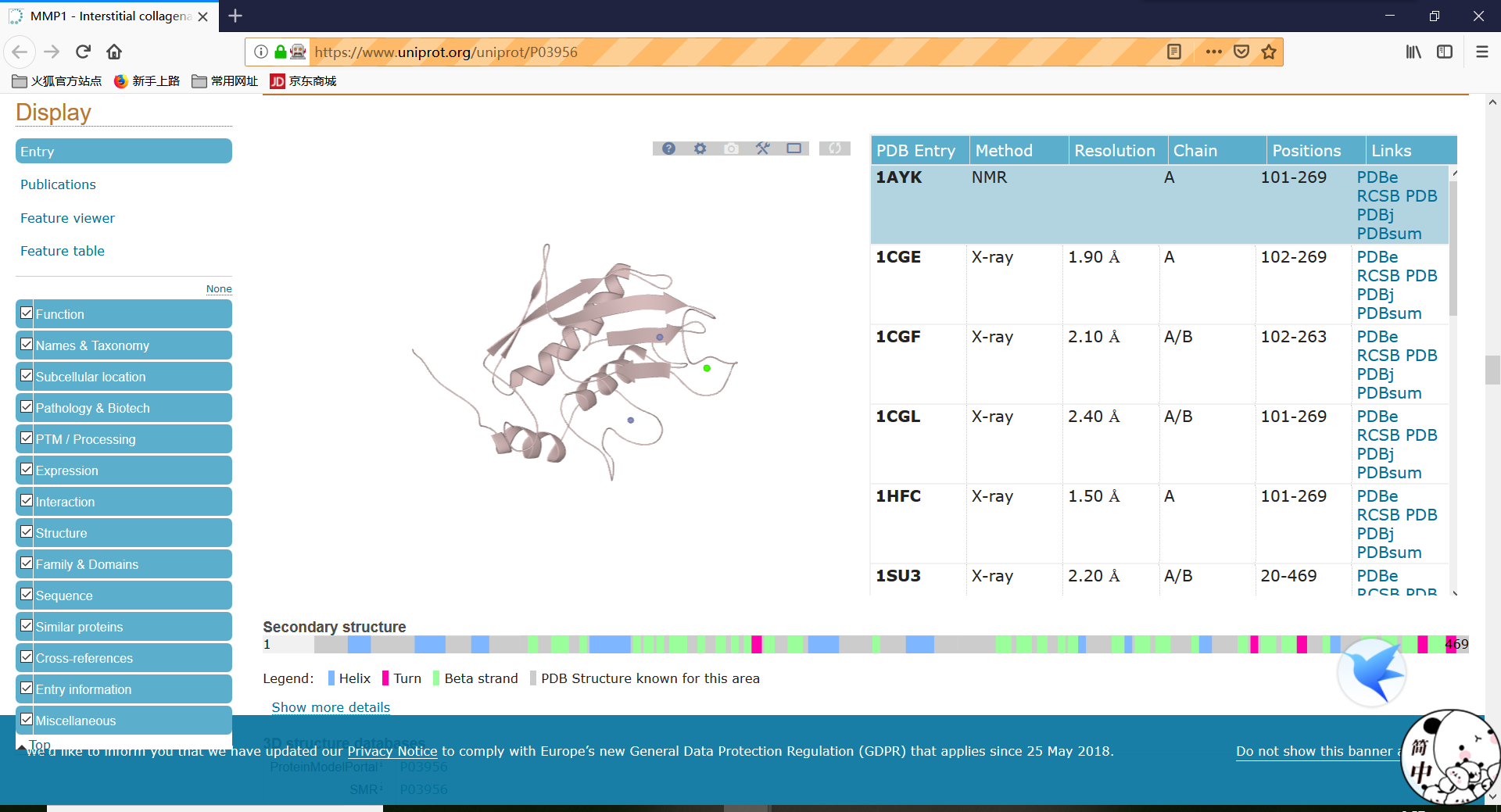
现在根据相关的资料，上述查询靶点中存在的问题通过查阅相关的文献了解到“同一个靶点对应的不同的PDB ID，实际上他们都是同一蛋白，只是在不同环境、不同分辨率、不同研究组或不同的其他条件下测得的结构”。这就是说，这里查询到的不同蛋白质是其实是同一种蛋白质，只是需要选择对应的PDB ID才好，现在的情况是，给定了靶点的名字，找到对应的靶点的数据。下面，对这个算法进行设计。下面是靶点蛋白的汇总。



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:

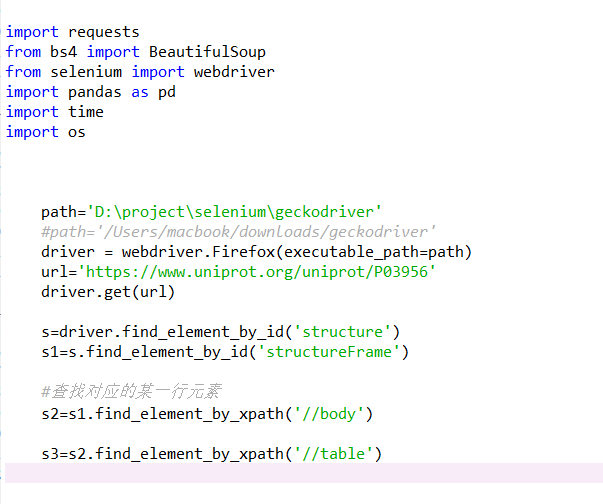


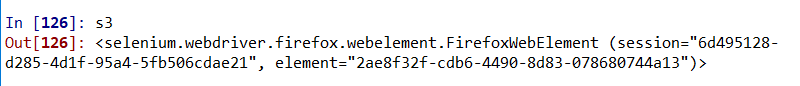
接下来，选择Human，然后再选择对应的Entry ,选择后点进去会进入到下面的页面，这个



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





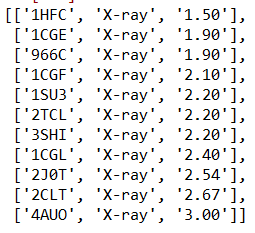
另外，需要注意的是这里有三个页面的代码需要处理，分别是当前的，前一个页面的代码，现在发现由于存在iframe框架，也就是说那个页面实际上是页面里边内嵌入的页面，这就导致不能直接进入到小页面中进行确定位置，所以处理的方式就是使用selenium的页面转换就是进入到子页面进行选择。最终定位到对应的页面，获取到页面中的PDB数据。然后需要对这些数据进筛选，筛选的算法是：

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。

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.