基于自动化技术的网络药理学研究

## 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。多维多层次关联分析

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。

### 摘要

我们提供了一个基于网络药理学的自动化在线中药分析系统，允许进行中药成分的准备、疾病靶点的准备、在线对接并给出对接的结果，系统集成上述三大模块。第一个模块根据提供的中草药的名称提供药品对应的小分子的名称以及ADME参数，对于非草药成分，系统提供本地成分上传控件，该模块对获取到成分会自动完成3D结构的获取。第二个模块会根据提供的疾病名称，自动搜索相关数据库完成疾病靶点、靶点蛋白的获取。第三部分根据前述得到的中药成分、疾病靶点结果进行在线批量对接，最终结果可以下载进行后续分析。本文提供了一个案例研究，演示了本系统如何发现中药成分的活性成分、疾病的关键靶点以及中药成分和靶点之间的作用关系，为中药网络药理学分析提供分析便利。现阶段的网络药理学分析，多专注于药物成分或者是疾病对应的靶点数据的获取，在线的分子对接平台也能进行分子对接。但是还没有一个集成的，可以从成分获取、靶点获取到对应的分子对接的平台。我们提供的系统集成这些功能并自动化实现。

### INTRODUCTION

网络药理学是融合了系统生物学（Systems biology）和多向药理学(Polypharmacology)提出的，

### 系统方法和界面

本系统基于网络药理学的思路，设计的本系统主要有三个部分，包括（1）中药成分的准备，包括草药的名称和非草药的数据excel格式上传2）输入疾病的名称，也可以根据用户的需要上传靶点蛋白3)进行对接，并查看对接结果，下载结果可以供用户下载使用。

### 中药成分准备

现阶段的中药物成分的查找很多是使用开源的数据库，例如西北农林科技大学开发的TCMSP（ http://lsp.nwsuaf.edu.cn/tcmsp.php）数据库、华东师范大学开发的中药综合数据库（TCMID，http://www.megabionet.org/tcmid/）、中国科学院上海有机化学研究所提供的中药化学成分数据库（http://202.127.145.134/scdb/）等相关中药信息数据库。但是药学研究人员在进行药物信息查找的时候仍旧面临着药物成分众多，不同的数据库之间存在着药物成分信息不同等诸多问题。同时，中药复方成分往往由多种草药构成，这给药物化合物成分的获取带来更大的麻烦。另外，不同的草药可能有相同的成分，这也需要进行后期的去重等相关操作。种种因素导致中药成分的准备过程复杂，耗时，往往需要花费大量的人力。另外，如果简单的根据限制药物的口服利用度（OB）和药物相似性（DL）来筛选药物的活性成分有会导致某些已经被验证的关键性活性成分缺失。

系统为了解决这些问题，并自动化处理这部分数据，采用了三种中药数据库融合的方式，把从这三种数据库中获取到的数据库进行自动化整合，首先TCMID和TCMSP数据库查询到的中药成分可以直接获取到PubChemID,虽然这两个数据库中的成分存在着名称不一致的问题，这是由于同一种成分可能会有很多的同名，但是PubChem是唯一的，一次为标准进行第一次数据库合并，上海有机所的中药成分数据库不存在PubChemID,同时成分的名称也存在异常，对此，将根据该数据库获取到的数据在PubChem数据库中的查找对应的PubChemID在于上面的数据进行合并。对于非草药数据，系统提供文本(excel)上传的

数据合并后，后台会根据得到的数据自动建立分别存储药物名称-药物成分，药物成分-药物的Pubchemid等两张表。然后根据ADME参数OB,DL进行药物的筛选。

筛选获取到中药成分后，需要使用到化合物成分的3D结构数据，每种成分的结构数据需要从NCBI(<https://www.ncbi.nlm.nih.gov/pccompound/>)等数据库中查找并自动下载。系统会把下载好的成分存储到服务器上。供后续的

### 疾病靶点准备

疾病的靶点选择是分子对接中重要的一部分，文献中介绍了GenBank、Genecard、Swiss Target Prediction、NCBI数据库获取到特定疾病对应的靶点。系统在调研的时候发现，NCBI数据库可以较为全面获取基因靶点，这里需要疾病的名称从NCBI数据库查找，在NCBI数据库，输入疾病的名称，限定查找Gene，然后限定物种为人类，获取相应地基因靶点名称。程序可以自动化实现疾病靶点的搜索。成分的靶点搜索完成后得到的是基因信息，还需要从UniProt数据库中获取到对应的UniprotID和PDB ID信息。根据成分的名称信息在UniProt数据库中进行查找的时候，可能出现不存在对应的靶点蛋白的情况，这种数据后续会筛掉；对于存在多个靶点蛋白的数据，由于数据量比较大，为了选择较为合理的靶点蛋白，同时达到快速选取靶点的目的，使用基于启发式算法的搜索策略选择，选择合适的靶点蛋白用作对接。

系统同时提供外部上传的方式进行疾病靶点的上传，这是为了解决特定靶点的分析而设计的。数据上传后可以自动进行数据的疾病的靶点蛋白,UniProtID以及相关的数据的准备。

### 在线对接

分子对接作为网络药理学分析中关键的一个步骤。传统的对接工具有Autodock,Discovery Studio,Autoduck Vina等对接工具，用于进行化合物小分子和靶点蛋白之间的对接，根据打分情况来判断小分子和蛋白质大分子之间的结合程度。上述的对接工具都是基于B-S架构，数据需要在本地进行处理。但是这些对接软件的打分虽然反应了结合程度，但是这些软件给出的打分都是自己制定的，分子对接的打分跨多很大，不能很好地从打分上看出对接的好坏。另外文献中提出了一些在线的工具，这些工具有的是从反向对接的方式来做，反向对接是从药物成分小分子出发反向分子对接技术是对分子对接的逆向思考,它以小分子化合物为探针,在具有三维结构的靶点数据库内进行分子对接,通过空间和能量匹配评价搜寻可能与之结合的生物大分子,进而预测理论上药物潜在的作用靶点，INVDOCK 平台、 PharmMapper都是基于这种方式进行对接的，对接结果会给出按照打分给出可能的作用靶点。但是，这些方法也能够进行化合物与靶点蛋白的结合程度的

判定。SystemDock是一款在线的分子对接平台，近年被开发出来后，大大方便了药理学的分析，该在线软件可以根据用户提供的化合物成分，疾病的靶点蛋白

进行在线分子对接，同时还可以给出对接的结果，提供了文件的下载等相关操作。

本系统集成了系统，根据准备好的重要成分和靶点自动导入该系统进行对接，最终把对接结果展现出来。但是该系统提供的数据仅仅包含靶点蛋白的PDBID、对接成分的PubChemCID等信息，但是在进行网络药理学分析的时候，往往需要更丰富的数据，例如靶点基因、UniProtID、Protein 等信息。系统在进行处理的时候会在SystemDock的基础上进行进一步的把前述相关信息提供给用户。对接结果可以供用户下载。

网络药理学研究需要准备大量的生物大分子和化合物小分子，但是当前进行这方面研究的时候，很多进行中药药理学方面的研究者需要进行大量的手动作业，导致数据获取进度缓慢，研究进展缓慢。针对这个痛点，本文提出了一种基于自动化技术的分子药理数据的准备方法。在给出中药药方以及其对应的病症的情况下，可以完成中药成分数据的获取同时构建对应的成分数据库。另外，同样还能完成病症靶点数据库的准备。该成功可以大大方便进行分子药理学的前期数据的准备，加速药理学的研究进展。同时，本文设计了一套新的靶点模块，用于根据

成分的准备主要分为两个模块，分别是生物大分子的模块和药品小分子的模块。这里，对这两个模块进行相关的设计与分析。

药品小分子的获取，首先使用到的数据库是TCMSP/TCMID这两个数据库，从这两个数据库获取到对应的小分子的名称数据，将这些名称数据存储在对应的文本中，然后分别访问这些小分子的名称，从对应的数据库中查找到对对应的SMILES格式的数据并且存储到对应的文件夹中，这就完成了对应的小分子的所有准备。

## 正文

### 现状

现阶段的中药物成分的查找很多是使用开源的数据库，例如西北农林科技大学开发的TCMSP（ http://lsp.nwsuaf.edu.cn/tcmsp.php）数据库、华东师范大学开发的中药综合数据库（TCMID，http://www.megabionet.org/tcmid/）、中国科学院上海有机化学研究所提供的中药化学成分数据库（http://202.127.145.134/scdb/）等相关中药信息数据库。但是药学研究人员在进行药物信息查找的时候仍旧面临着药物成分众多，不同的数据库之间存在着药物成分信息不同等诸多问题。同时，中药复方成分往往由多种草药构成，这给药物化合物成分的获取带来更大的麻烦。另外，不同的草药可能有相同的成分，这也需要进行后期的去重等相关操作。种种因素导致中药成分的准备过程复杂，耗时，往往需要花费大量的人力。另外，如果简单的根据限制药物的口服利用度（OB）和药物相似性（DL）来筛选药物的活性成分有会导致某些已经被验证的关键性活性成分缺失。

在获取到中药成分后，根部不同的对接方法，需要使用到化合物成分的3D或者是SMILES结构数据，每种成分的结构数据需要从NCBI

(<https://www.ncbi.nlm.nih.gov/pccompound/>)等数据库中查找并下载，由于药物成分众多，这同样导致成分的结果数据获取困难，需要花费大量的人力物力。

靶点选择也是重要的一个部分，这里需要根据疾病的名称从NCBI数据库查找，在NCBI数据库，输入疾病的名称，限定查找基因，然后限定物种为人类，获取相应地基因靶点名称。但是，一般情况下基因的数据很多，同样需要花费大量的时间来处理。或者是使用DisGeNET数据库使用其基因查找功能，查找到对应的基因靶点。在获取到基因靶点名称后，还需要查询其对应的靶点蛋白，使用Uniprot数据库的UniProtKB搜索功能，输入对应的基因靶点名称，获取对应的靶点蛋白、Uniprot ID，同时还需要筛选靶点蛋白的PDB ID。一种疾病的基因靶点众多，另外又由于Uniprot本身设计的问题，查找基因靶点的过程需要大量重复性的工作，同时稍有不慎还会导致错误。另外，如果不查找到所有的靶点信息，仅仅根据预定的某些规则选取部分基因靶点，也可能导致后续数据的缺失。

分子对接的数据准备是进行分子对接中关键的操作，对接本身也同样十分重要，Systems dock Web Site是一种基于网络药理学的预测和分析的网络服务器，它允许对接模拟和分子途径图，用于综合表征配体选择性和解释复杂分子网络上的配体作用。但是，该在线工具在使用的过程中，存在着诸多的问题，例如，每次的化合物成分不能超过5个，每次都需要进行手动上传靶点蛋白的PDB ID和成分的结构，过程较为复杂，后期的数据处理也较为麻烦，导致研究进程缓慢。

### 研究进展

为了解决中药网络药理学研究中出现的数据收集，分子对接的问题，设计了一个自动化网络药理学平台，从中药化合物分子的获取，疾病靶点的获得，到最后对接的过程，以及对接数据的分析，组合成一个系统，可以一次性完成这些任务，大大加快了网络药理学研究的进程，可以快速的分析中药成分与疾病靶点之间的关系，在较短的时间内，构建活性成分-靶点、靶点-靶点、成分-靶点-疾病、成分-靶点-通路网络来阐述药物的作用机制，快速进行中药复方作用机制的多成分、多靶点研究。

下面是软件的架构图，网站由后台和前端两个部分组成，后台有三个部分组成，这里首先详细介绍后台的组成。本平台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TCMID、TCMSP两个数据库中查找对应的药品化合物成分数据。两个数据库的数据的数据是根据手动查询药物的过程，使用自动化框架selenium+python进行开发，从药品名称的输入，到相关的定位，到最后查询结果的输出，输出会以excel的格式输出到本地。化合物3D文件的下载，是从接下来是成分3D结构的下载，这里使用了PubMed数据库进行，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。实验使用selenium+python模拟下载3D结构的操作，使用循环处理的方式，将查询的药品化合物成分的3D结构下载到本地后续处理。这里设计主要是为了进行草药（herb）的化学成分的获取，如果是非草药的数据，本云平台还可以满足上传scv或者是excel的方式进行非草药数据的获取。

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，然后使用selenium+python设计的程度自动化获取查询到的靶点信息并存储到本地。对于有些数据是不能直接找到靶点的，需要使用另外的数据库来处理，云平台后台使用 NCBI数据库进行查找，查找需要在数据库的查询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基因靶点，云平台自动模拟该部分操作，并将全部数据爬取存储在对应的excel文件中。后续所以需要在基因靶点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蛋白。靶点蛋白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的id信息，这个也需要使用UniProt的靶点蛋白页面进行获取。这一步同样可以使用本云平台的设计进行自动化的获取。

前述部分获取完成后，接下来进行的是云平台的对接，云平台后台整合了

System Dock Web Site在线对接系统，程序完全模拟人工的对接操作，实现从化合物小分子、蛋白质靶点、对接、以及给出对应的结果。

### 实验验证

基于本在线网站，这里进行案例研究，这里对散结镇痛胶囊治疗子宫肌腺症进行案例研究。

散结镇痛胶囊由四种中药组成，分别是龙血竭、三七、浙贝母、薏苡仁，其中除了龙血竭不是草药之外，其他的都是草药，所以可以直接输入后三者的名称，使用云平台进行成分的获取，由于龙血竭不是草药，不能从相关中药数据库中获取，所以需要自己进行文献挖掘获取到对应的数据，从文献中查找到的龙血竭的成分信息以excel的方式上传到服务器。

上述信息完成后，会得到四种中药材总共 种成分，这个过程很快完成了。

接下来使用云平台输入疾病名称adenomyosis，endometriosis，获取对应的因靶点，该过程需要花费一定的时间来完成，但是很快程序会得到给定全部的靶点数据，从结果看总共是个基因靶点个数是 个。上面两部完成后，使用云平台的对接公共，点击对接按钮，进入对接页面，使用本案例选取到的中药成分和靶点蛋白进行对接，然后最后会得到对接的结果，对接结果会以打分的形式给出，最后下载下来对应的结果得到所有的成分-靶点打分信息，下载下来后