基于自动化技术的分子对接研究

# 摘要

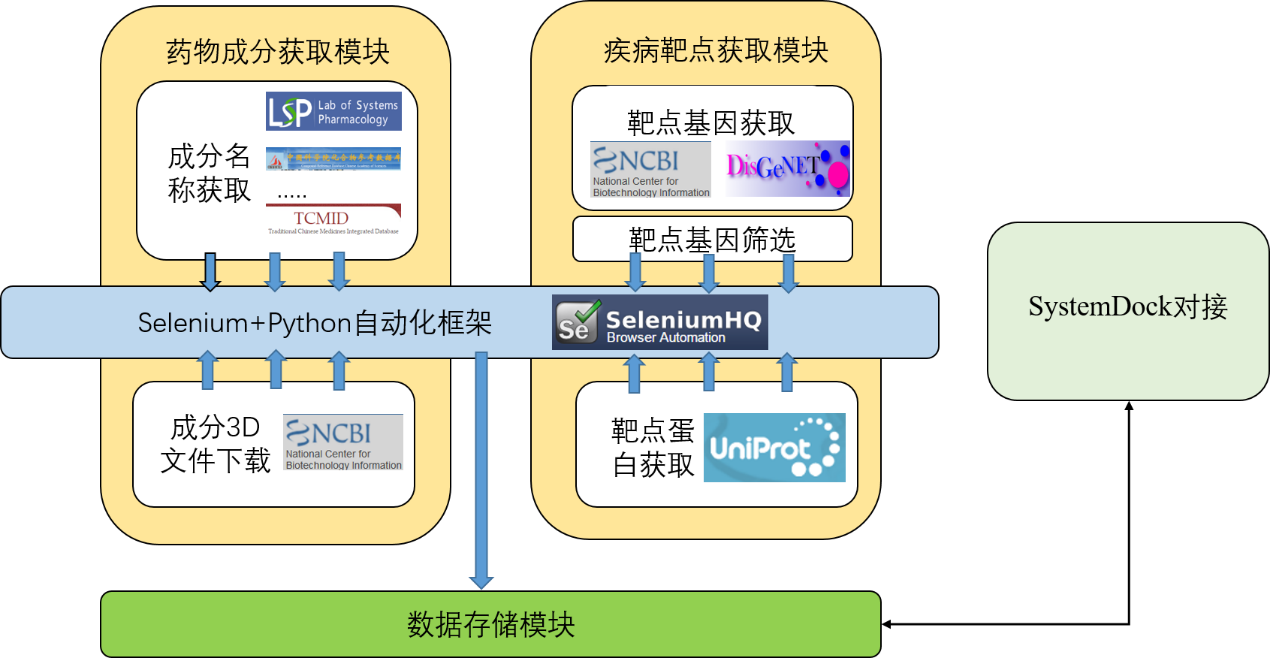
我们提供了一个基于网络药理学的自动化在线中药分析系统，允许进行中药成分的准备、疾病靶点的准备、在线对接并给出对接的结果，系统集成上述三大模块。第一个模块根据提供的中草药的名称提供药品对应的小分子的名称以及ADME参数，对于非草药成分，系统提供本地成分上传控件，该模块对获取到成分会自动完成3D结构的获取。第二个模块会根据提供的疾病名称，自动搜索相关数据库完成疾病靶点、靶点蛋白的获取。第三部分根据前述得到的中药成分、疾病靶点结果进行在线批量对接，最终结果可以下载进行后续分析。本文提供了一个案例研究，演示了本系统如何发现中药成分的活性成分、疾病的关键靶点以及中药成分和靶点之间的作用关系，为中药网络药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。最后系统给出了案例研究，研究散结镇痛胶囊治疗子宫肌腺症的治疗。

# 背景

中药是中国传统医学中重要的，现代中药的研究进入到成分-基因-靶点层面，中药的起作用的特点是多成分，多靶点。网络药理学是研究现代中药关键的技术。分子对接技术是网络药理学中关键的一个部分。分子对接是把药物化合物和疾病对应的靶点按照某种方式进行结合，然后按照结合程度进行打分的一种方法。根据打分的结果判断成分和疾病对应的靶点之间是否起作用。现阶段的中药成分获取是从TCMSP、TCMID、上海有机所等关键的中药数据库中获取的成分，传统的方式需要手动查找，多库之间还存在数据的异构性等问题，为了处理这些数据，设计了专门的程序来把不同库里的数据整合起来，然后把这些数据梳理成合适的格式。但是现阶段还没有能够一个系统可以完全获得某种成分药物成分，所有的成分都需要手动查询并全部并需要手动文本挖掘成分。疾病靶点的获取也是一个重要的部分，众多研究者在进行靶点数据获取的时候，从NCBI、Genecard等数据库中手动查找，往往需要耗费大量的时间，对接使用的靶点PDB ID也需要特定的筛选，本系统设计的可以根据疾病的名称直接获取到对应的全部的靶点信息。网络药理学最关键的一个步骤就是对接，现阶段的存在很多的对接软件，autodock, Discovery Studio都可以进行在线对接，在线的对接工具SystemDock是近年开发的可以快速对接的平台，软件基于Autodock Vina开发，重新设计了打分函数，可以根据打分评价分子和靶点结合的好坏。系统集成了该平台，可以批量进行对接操作，同时可以导出对接数据。本文会从系统方法和界面、软件模块和算法、案例研究等三个方面进行系统的介绍。

# 系统设计

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。下面是本文的结构图，结构图展示了整个数据的处理流程。



## 中药成分模块

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，一次为标准进行数据合并，对于非草药数据，系统提供文本(excel)上传。

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2D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

## 疾病靶点模块

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。

## 在线对接模块

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。

# 设计细节

## 中药成分模块

中药成分的获取是该系统较为复杂的部分，现阶段由于中药成分数据库很多，但是这些中药成分数据之间存在异构性，不同数据来源于不同的数据库，同时这里根据不同的模块的数据处理设计了不同的方法。中药成分的获取，数据需要从TCMSP、TCMID或者是上海有机所的数据库中查找，但是，单纯的数据查找不能全面找到对应的信息。例如从TCMSP和TCMID以及上海有机所的数据库中都不能获取到成分的2D/3D结构信息，该信息只能从PubChem等数据库中下载，所以前三者的数据需要进行统一。由于系统需要的数据最关键的是药物的名称、成分名称、成分的PubChemCID,成分的2D/3D结构，成分的ADME参数OB,DL，所以每个数据库都需要获取到该部分数据，对于获取不到全部的，程序会有所取舍。但是实际上这三个部分获取数据的方式在基本模式设计上是一致的。

获取数据库中数据设计的总体思路如下所示，首先需要根据名称获取到成分的主页面url，然后解析得到每个成分对应的url，获取靶点后还需要在PubChem数据库中按照搜索到的PubChemCID获取到对应的2D结构并下载。

|  |
| --- |
| \_D:\MarinaJacks\project\reptilian\medicine\docs\研究文档\分子对接数据\TCMID.png |
| 中药信息数据获取流程设计 |

在最终获取到成分2D结构的同时，需要把药物名称、化合物名称、PubChemCID等信息存储到数据库中，2D结构的命名与PubChemCID保持一致。

对于比较特别的需要非草药成分，系统需要根据上传文件的成分名称信息从NCBI Compound数据库中查找成分的PubChemCID,并下载对应的成分的2D文件。相关的数据会存储到对应的数据库中。

最后，完成中药成分的准备。完成准备后的数据是包括中药成分数据库，中药2D结构数据。

## 靶点成分模块

该模块可以自动化实现根据给定的疾病自动从NCBI数据库获取到靶点基因的Gene,然后需要在UniProtKB数据库查询对应的UniProtID,顺便获取到对应的靶点蛋白名称，最后获取到对应的PDBID数据。在进行PDBID获取的时候，每种靶点蛋白有多种结构，需要对结构进行进一步的筛选，传统的筛选方式是根据靶点蛋白的结构分辨率进行排序，然后筛选出小分辨率的靶点蛋白PDBID，获取分辨率较小的靶点蛋白作为合适的靶点蛋白，系统会按照这个方式把靶点的名称、靶点蛋白名称、UniProtID、PDB ID都存储到对应的数据库中。完成疾病靶点的准备。对于该部分需要用户自己提供上述靶点名称、靶点蛋白名称、UniProtID、PDB ID等信息。

## 在线对接模块

在线对接的后台使用的是systemsDock Web Site，该模块底层使用的是AutoDock Vina分子对接工具，同时该平台在AutoDock Vina的打分基础上，使用docK-IN打分算法，对原始的打分进行了重新计算，使得打分结果能直接描述结合强度。一般打分值在4.25 以上认为是有一定的结合活性，大于5.0 的说明分子与靶点有较好的结合活性，大于7.0 则说明分子与靶点的结合构型具有强烈的活性。本系统封装了该在线模块，可以直接在前面的步骤基础上，给出对接结果。原始在线软件每次仅能提供5个化合物成分对接，本系统可以一次性完成所有成分的对接。另外，原始对接结果仅仅给出了对接的化合物成分的编号和靶点的PDBID信息，本系统在该功能的基础上，化合物成分名称，所属中药，靶点的其余信息，例如UniprotID,Protein name，gene name等也一并给出，最终提供下载。

系统封装的4个步骤，分别是SystemDock的靶点PDBID上传，中药成分的化合物2D/3D结构上传，对接以及对接结果的下载。

对接完成后，系统提供了化合物成分信息，靶点成分信息，对接结果信息等3个部分的下载。

# 案例研究

基于本系统，本文做了一个案例研究，本案例研究散结镇痛胶囊治疗子宫肌腺症的治疗，分析成分对疾病对应的靶点的作用情况。

首先，使用中药成分模块，首先根据资料查阅到散结镇痛胶囊的成分为龙血竭、三七、浙贝母、薏苡仁，其中三七、浙贝母、薏苡仁都是草药，这个可以直接使用系统的关键字查找功能来获取化合物成分，龙血竭不是草药，这种药物的成分就需要从文献中查找，然后使用系统的文本上传功能来查找成分。

然后根据治疗的疾病，子宫肌腺症，其对应的英文为adenomyosis，endometriosis,在系统的疾病靶点模块输入这种疾病对应的疾病名称。

前两步完成后，执行对应的对接操作，等待对接完成。然后分析给出的下载数据。下载数据提供的3个数据表，分别时中药成分数据表，靶点信息表和对接结果信息表。根据下载的3个表，可以得到对应的中药成分信息，靶点信息以及对接结果信息。

根据中药成分信息表可知，这四种中药材对应的成分去重之后得到种242种成分，其中 根据 有

根据得到的靶点成分信息，得到该疾病对应的靶点有56个，如下所示。

|  |  |  |  |
| --- | --- | --- | --- |
| **UniProtID** | **Protein** | **Gene** | **PDBID** |
| P08253 | 72 kDa type IV collagenase | MMP2 | 3AYU |
| P15692 | Vascular endothelial growth factor A | VEGFA | 1MKK |
| P03372 | Estrogen receptor | ESR1 | 3CBP |
| P35222 | Catenin beta-1 | CTNNB1 | 3FQN |
| P06401 | Progesterone receptor | PGR | 1SQN |
| P14780 | Matrix metalloproteinase-9 | MMP9 | 6ESM |
| P10451 | Osteopontin | SPP1 | 3CXD |
| P01138 | Beta-nerve growth factor | NGF | 1WWW |
| Q13651 | Interleukin-10 receptor subunit alpha | IL10RA | 1Y6K |
| P35354 | Prostaglandin G/H synthase 2 | PTGS2 | 5F19 |
| P21964 | Catechol O-methyltransferase | COMT | 3BWY |
| Q92731 | Estrogen receptor beta | ESR2 | 3OLL |
| P05106 | Integrin beta-3 | ITGB3 | 1MIZ |
| P03956 | Interstitial collagenase | MMP1 | 1HFC |
| P16035 | Metalloproteinase inhibitor 2 | TIMP2 | 1BR9 |
| P04629 | High affinity nerve growth factor receptor | NTRK1 | 5JFW |
| P08138 | Tumor necrosis factor receptor superfamily member 16 | NGFR | 3EWV |
| P08887 | Interleukin-6 receptor subunit alpha | IL6R | 1N26 |
| Q16665 | Hypoxia-inducible factor 1-alpha | HIF1A | 4H6J |
| P19838 | Nuclear factor NF-kappa-B p105 subunit | NFKB1 | 1SVC |
| P42771 | Cyclin-dependent kinase inhibitor 2A | CDKN2A | 1BI7 |
| P29474 | Nitric oxide synthase, endothelial | NOS3 | 4D1P |
| O00206 | Toll-like receptor 4 | TLR4 | 2Z62 |
| P60484 | Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN | PTEN | 4O1V |
| P10145 | Interleukin-8 | CXCL8 | 4XDX |
| P37231 | Peroxisome proliferator-activated receptor gamma | PPARG | 3U9Q |
| P09488 | Glutathione S-transferase Mu 1 | GSTM1 | 1XW6 |
| P42345 | Serine/threonine-protein kinase mTOR | MTOR | 4DRI |
| P24385 | G1/S-specific cyclin-D1 | CCND1 | 2W96 |
| P38936 | Cyclin-dependent kinase inhibitor 1 | CDKN1A | 5E0U |
| P13500 | C-C motif chemokine 2 | CCL2 | 1DOK |
| O15392 | Baculoviral IAP repeat-containing protein 5 | BIRC5 | 2QFA |
| P46531 | Neurogenic locus notch homolog protein 1 | NOTCH1 | 5L0R |
| O95256 | Interleukin-18 receptor accessory protein | IL18RAP | 3WO4 |
| P17693 | HLA class I histocompatibility antigen, alpha chain G | HLA-G | 3KYO |
| P06400 | Retinoblastoma-associated protein | RB1 | 2R7G |
| Q15109 | Advanced glycosylation end product-specific receptor | AGER | 5D7F |
| P04798 | Cytochrome P450 1A1 | CYP1A1 | 4I8V |
| P09038 | Fibroblast growth factor 2 | FGF2 | 4OEE |
| P08254 | Stromelysin-1 | MMP3 | 1HY7 |
| P01033 | Metalloproteinase inhibitor 1 | TIMP1 | 3V96 |
| P05093 | Steroid 17-alpha-hydroxylase/17,20 lyase | CYP17A1 | 5IRQ |
| P09237 | Matrilysin | MMP7 | 2Y6D |
| O95863 | Zinc finger protein SNAI1 | SNAI1 | 4QLI |
| Q99497 | Protein/nucleic acid deglycase DJ-1 | PARK7 | 2RK3 |
| Q9Y6K1 | DNA (cytosine-5)-methyltransferase 3A | DNMT3A | 4QBS |
| P13591 | Neural cell adhesion molecule 1 | NCAM1 | 2HAZ |
| Q13464 | Rho-associated protein kinase 1 | ROCK1 | 3V8S |
| Q9UBC3 | DNA (cytosine-5)-methyltransferase 3B | DNMT3B | 5NRR |
| Q969J5 | Interleukin-22 receptor subunit alpha-2 | IL22RA2 | 3G9V |
| O00622 | Protein CYR61 | CYR61 | 4D0Z |
| P05230 | Fibroblast growth factor 1 | FGF1 | 1RG8 |
| P42702 | Leukemia inhibitory factor receptor | LIFR | 3E0G |
| Q9NZH6 | Interleukin-37 | IL37 | 5HN1 |
| O96013 | Serine/threonine-protein kinase PAK 4 | PAK4 | 2J0I |
| P49757 | Protein numb homolog | NUMB | 5NJJ |

根据对接的打分结果，