第3章　网络药理学常用数据库

本章导读：

网络药理学是在生物学大数据和人工智能背景下产生的，数据库对于网络药理学研究至关重要。历朝历代的中医药古籍文献有包括海量的方剂。现代研究对很多方剂或者中药材进行了成分分离和分析，以及现代分子药理研究。当前，已经系统整理了一些中医药领域内的重要数据库，这些数据库多从中药复方或药材的组成化合物出发，通过药物潜在的靶点，利用网络药理学的手段建立中药与疾病或者证候之间的关联。这些数据库为认识中药治疗疾病的机制，以及加深对中医药理论的理解提供了值得发掘和进一步验证的资源。

除了中医药数据库之外，网络药理学研究离不开一些国际上重要的公共数据库。例如，药物和化学数据库为我们认识中药成分等天然产物的理化性质、生物活性、作用靶标、成药性等方面提供了数据。同时这些数据库收录的FDA批准的已上市药物的信息，也为药物信息学的研究提供了金标准。此外，OMIM、HPO(Human Phenotype Ontology)、DisGenet等疾病数据库为探究疾病相关基因及疾病发生机制提供了丰富可靠的注释信息。而包括STRING在内的蛋白质-蛋白质相互作用数据库则为建立药物与疾病之间的关联，构建药物干预疾病网络提供了丰富数据。

因此，本章将对网络药理学常用的数据库进行详细介绍，主要从中医药数据库、化学及药物数据库、疾病和蛋白质相互作用数据库三方面展开。

3.1　网络药理学常用中医药数据库

中药复方由多种中草药构成，每种中草药还包含多个活性成分，因此也就导致了中药的作用靶点是广泛的。但正是因为中药“多组分、多靶点、多通路”的作用机制，使它能够有效治疗包括癌症和糖尿病等在内的复杂疾病。基于中药的以上特点，利用网络药理学的思想研究中药的作用机制可能是一种有效的方式。网络药理学研究中涉及中药成分、靶标、通路、表型、证候、疾病等多种实体。ETCM(The Encyclopedia of Traditional Chinese Medicine)[1]和TCMID(Traditional Chinese Medicines Integrated Database)[2,3]等数据库注重于中药相关的化学成分、作用靶标等数据的收集。而SymMap(Symptom Mapping)[4]、TCMGeneDIT[5]则关注中药实体之间的关联，其中SymMap收录并评价了中医症状、西医症状与中药成分、靶点之前的关联。而TCMGeneDIT则通过文献挖掘来构建以及评价中药、基因、疾病之间的关系。TCMSP(Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform)[6]以及BATMAN-TCM(Bioinformatics Analysis Tool for Molecular Mechanism of Traditional Chinese Medicine)[7]则以基于成分的靶标预测和网络分析为核心。这些数据库均为中药机制研究提供了重要的数据。本章对近年来的网络药理学相关数据库及分析平台进行简要的介绍，旨在通过本章熟悉网络药理学相关的中医药数据库，了解可以用于网络药理学研究的中医药数据资源和

数据平台。对于每一个中医药数据库，本章将从数据库简介、数据库结构、主要功能、数据库特性等方面进行介绍。

3.1.1　ETCM：中医药百科全书

ETCM[1]（The Encyclopedia of Traditional Chinese Medicine）是 2018 年由中国中医科学院中药研究所许海玉团队与北京大学药学院天然药物和仿生药物国家重点实验室刘振明教授等共同设计开发的一个中药综合资源数据库。ETCM 的主要功能包括：①提供关于常用中草药、中药复方及其所含成分的全面且标准化的信息，为用户获取关于中药及方剂的全面信息提供便利资源；②根据中药成分和已知药物之间的化学指纹相似性，进行中药成分的靶标预测；③系统分析功能，用户能够在网站内建立网络来探索中药、复方、成分、基因靶点和相关作用途径或疾病之间的关系。ETCM 基于网络药理学策略，旨在阐明中药与靶标和现代疾病之间的潜在联系，揭示中药的作用机制，为促进中医药相关基础研究、临床应用和药物开发提供重要资源。

1. 数据结构

ETCM 中汇集了 403 味中药（产地、性味归经、适应症、所含成分、质量控制标准等）、3962 个中药复方（名称、剂型、组成、适用症、所含成分等）、7274 种中药化学成分、2266 种有效或预测的药物靶标以及 3027 种相关疾病（如图 3-1 所示）。又进一步将中药按照药味（酸、苦、甘、辛、咸）、药性（寒、热、温、凉、平）、归经（肺经和肝经等）进行分类，通过单击上述每个类别的饼图，用户即可获得属于每个类别的中草药的完整列表。每种中草药的详细信息可以通过单击其中文或拼音名称来检索，包括产地、最佳采收时间、性味、归经、适应症和所含化学成分，每味药的图片及其在中国的产地分布、质量控制标准。中药的信息页面中还提供了包含该味药的所有复方名称，单击每个复方名可以直接链接到复方的信息页面。由特定成分、中药、复方或与特定疾病相关的基因所富集的**基因本体（Gene Ontology，GO）**或通路也包含在 ETCM 中（如图 3-1 所示）。

2. 功能介绍

1）中药成分靶标预测

该库使用 MedChem Studio（3.0 版）来预测中药成分的潜在靶标，MedChem Studio 是一种药物相似性搜索工具，用于查找与中药成分具有高度结构相似性（Tanimoto＞0.8）的已知药物，从而进行靶标预测。Tanimoto 的值限定在[0，1]的范围内，其中“0”表示成分和已知药物之间结构完全不同，“1”表示两种成分具有相同的结构。针对用户输入的中药成分，MedChem Studio 经过筛选后，得到 Tanimoto＞0.8 的候选靶标列表。候选药物靶标的生理功能和参与途径从 Gene Ontology 和 KEGG 数据库中获取。

2）网络分析

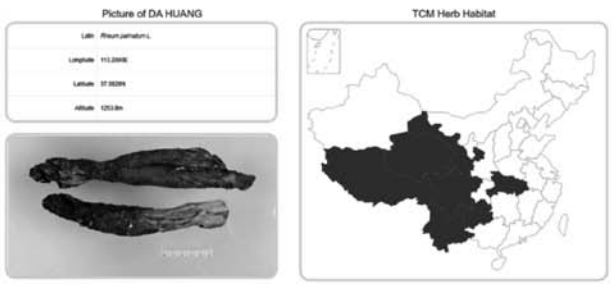
为了更好地说明成分、中药、复方、靶标、涉及靶标的通路和疾病之间的关系，ETCM 提供了系统分析功能，允许用户在上述两个或多个项目之间建立网络。通过输入查询项并选择一个或多个类别，用户能够在系统内构建中药—成分—靶标、复方—中药—通路、复方—中药—靶标—疾病以及其他网络，如图 3-2 所示。还可以在其中标记或修改网络的节点和边缘，以方便进一步的研究。

3）基于化合物 ADMET 预测评估成分的类药性

为了评估每种成分的类药性，ETCM 中还提供了基于 Pipeline Pilot 平台的 ADMET（药物的吸收、分配、代谢、排泄和毒性，Absorption，Distribution，Metabolism，Excretion，Toxicity）模块计算得到的各成分的药代动力学参数，包括水溶性、血脑障碍渗透性、CYP450 2D6 抑制率、肝毒性、人体肠内吸收和

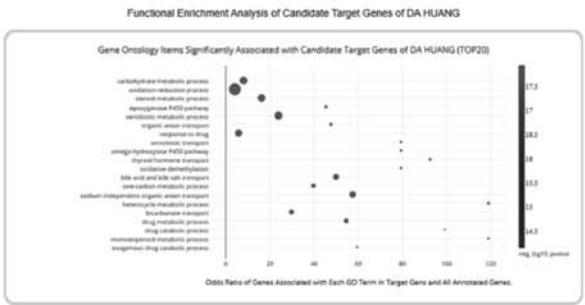
(a) ETCM功能概述

(b) ETCM中草药部分，单击饼图的任意比例部分，将在下面的列表中显示此类中全部的中草药

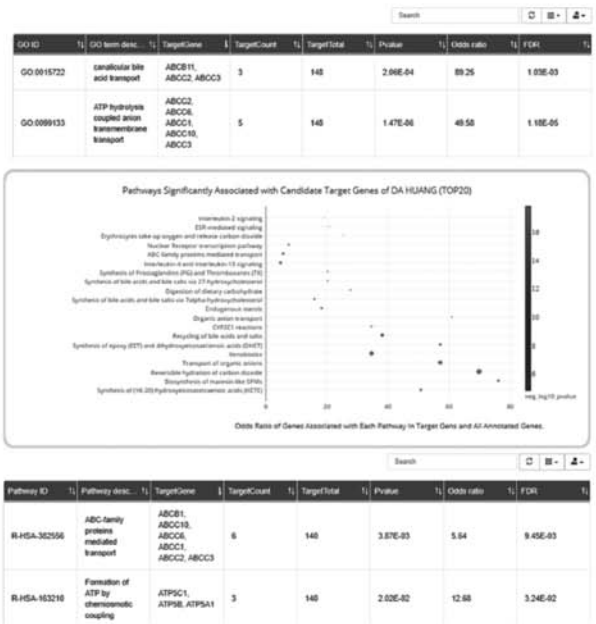


(c) 草本植物大黄的图片和产地地图

(d) 大黄的质量控制标准

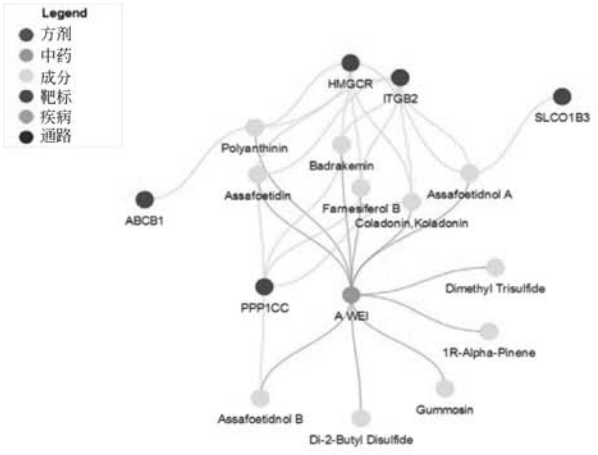


(e) 大黄的预测靶基因的富集和GO表



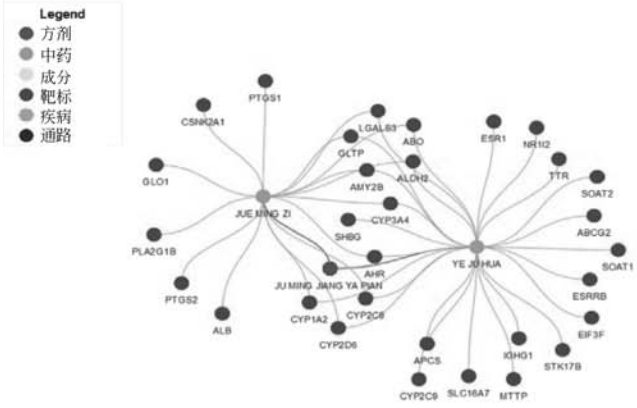
(f) 大黄的预测靶基因的富集通路

图 3-1 ETCM 的主要功能说明

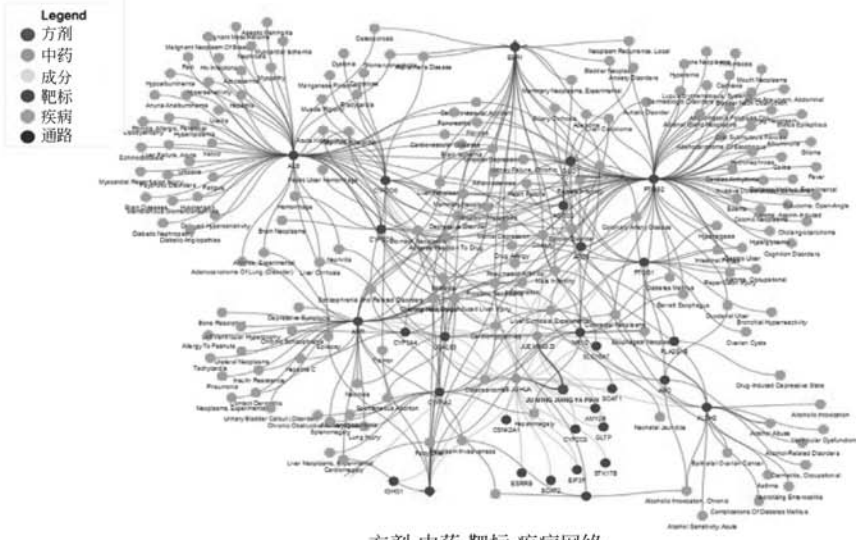


中药-成分-靶标网络

方剂-中药-通路网络



方剂-中药-靶标网络



方剂-中药-靶标-疾病网络

图 3-2　ETCM 系统分析功能构建的各种网络示例

血浆蛋白结合率等。并运用 QED(Quantitative Estimate of Drug-Likeness)来定量评估成分的类药性，QED 的取值范围为(0,1)，QED 值为 0 表示该化合物所有的性质都不利于成药；而 QED 值为 1，说明该化合物的成药性极好。有研究表明，在药物开发中有吸引力成分的平均 QED 值为 0.67，无吸引力成分的平均 QED 值为 0.49。据此，ETCM 将其中收集的所有 7274 种中药成分按照其 QED 值(即成药性)分为三组：好(QED＞0.67)、中等(0.49≤QED≤0.67)和弱(QED＜0.49)，为后续的成药研究提供一定的依据。

3. 特性

(1) 2015 版《中国药典》中提供的适应症不同于现代疾病，因此 ETCM 尝试使用中药成分和现代疾病间的基因关系来建立中药适应症和现代疾病之间的联系。

(2) 网络分析：为了更好地说明成分、中药、复方、靶标、涉及基因的通路和疾病之间的关系，ETCM 使用基于动态浏览器的可视化库 vis.js(4.21.0)网络模块，用户可以构建中药、复方、靶标与疾病之间多级交互的网络。

(3) 与其他中医药相关数据库相比，ETCM 增加了新的模块和功能，包括中药的产地分布图、中草药的图片、中药及复方的质量控制标准、指标性成分的定量信息、成分的 ADME(药代动力学)参数、药物相似性评价、ChEMBL 和 PubChem 数据库的链接、网络构建和分析等。

3.1.2 SymMap：关注证候关联的中医药整合数据库

SymMap[4](Symptom Mapping)是一个注重证候关联的中医药整合数据库。在该数据库中收录了中医症状、中草药、西医症状关联的疾病、中草药成分、药物靶点，而这六种类型的实体之间的关联也构成了一个异质网络。SymMap 通过这种方式将中国传统医学与现代医学从表型和分子层面都加以关联。在 SymMap 中，六种类型实体之间的关联关系都基于统计检验加以评价和打分，药物学家能够根据重要程度进行筛选进而指导药物发现。

1. 数据结构

SymMap 的六大类实体库包括 1717 种中医症状，961 种西医症状，499 种中草药，19595 种药物成分，4302 种药物靶标和 5235 种疾病。SymMap 的六类实体之间的直接关联包括 6638 种中草药-中医症状关联，2978 种中医症状-西医症状关联，48372 种中草药-药物成分关联，12107 种西医症状-疾病关联，29370 种药物成分-药物靶标关联和 7256 种基因-疾病关联(如图 3-3 所示)。例如，在中医症状-中草药关联中，每种中草药平均与 13.30 种中医症状相关，每种中医症状平均与 3.87 种中草药相关。在 SymMap 提供的中医症状-西医症状关联集合中，每种中医症状与 1.74 种西医症状相关，每种西医症状与 3.13 种中医症状相关。

2. 功能介绍

1) 检索方式

用户可以通过 SymMap 的网站页面浏览、搜索和下载其六个部分和互相关系。可以单击主页上的搜索按钮，并在搜索页面输入检索条目完成搜索。SymMap 的每个部分可以提供一个包含不同关键词的多类型的搜索框。例如，在搜索特定的西医症状时，允许使用三种不同的关键词，包括症状名称、其他

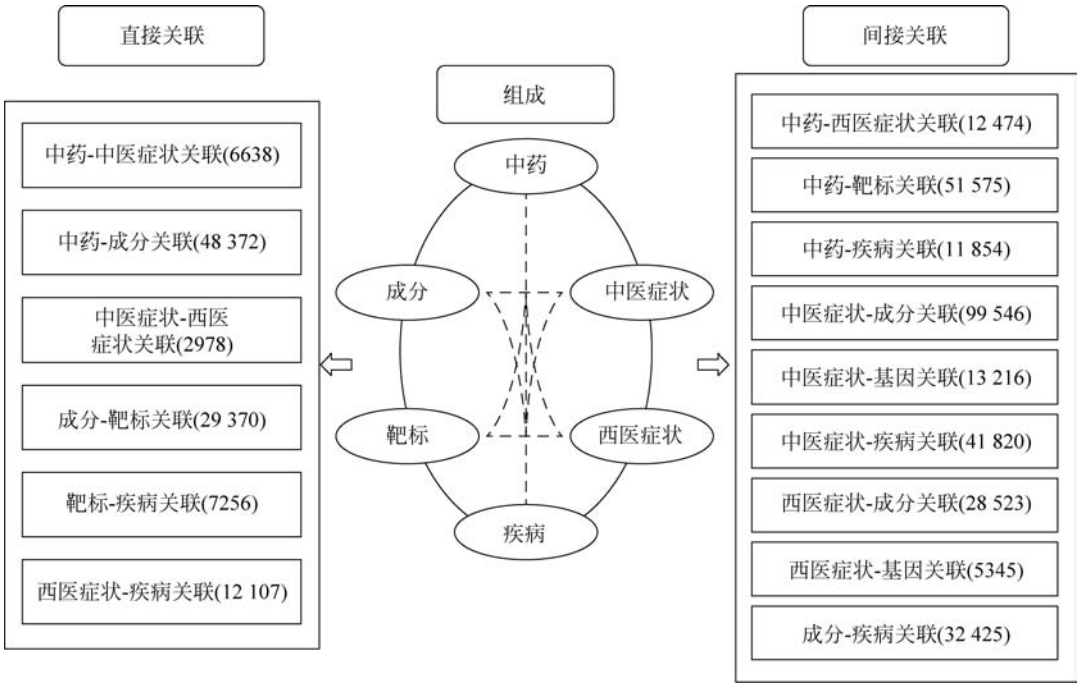


图 3-3　SymMap 数据构架图

图片中间是 SymMap 中包括的六类实体集。六类实体之间的直接连边表示六种直接的实体关联，它们的名称和数量列在左侧。而六类实体之间的九种间接实体关联则被列在右侧

公共数据库中收录的症状 ID 以及症状的同义词。用户还可以下载 SymMap 的检索结果。另外，在输入检索条目后，SymMap 的自动搜索功能会提供相似的词条供用户选择，进而完成 SymMap 的搜索。

2）SymMap 的检索结果

SymMap 搜索结束后，符合条件的条目在搜索界面下方的一个总结表中展示，第一列为 SymMap ID。用户可以单击 SymMap ID 的超链接获得细节信息。在细节界面，SymMap 提供详细的描述信息，并通过检索条目与其他五个部分关联的网络可视化图像以及表格。此外，在六个部分中所有 item 的列表可以在浏览界面浏览，并且所有列表都可以通过网页下载。

在浏览或搜索 SymMap 后，用户可以单击每个特定条目的 SymMap ID 进入细节界面，该界面会提供包含条目的摘要、六个部分间可视化相互关系的网络面板，以及展示搜索条目与其他五个实体之间的关联列表。

3）摘要面板

摘要面板展示检索项的摘要信息，如图 3-4 所示。SymMap 提供三种信息：①名称和基因符号（Gene Symbol）；②解释信息（定义与分类）；③其他数据库中的外部链接，可以直接单击进入其数据库。

4）网络面板

网络面板提供检索项与其他实体关联网络的可视化，如图 3-5 所示。网络中的节点根据其类型被标以不同颜色，并放置在不同位置。节点的大小由其在网络中的连接度决定。当用户将鼠标悬置在某个节点上方，节点会变大，且与其相关的关联将高亮。此外该节点的名称在气球框中显示。

图片中的每个节点由超链接连到其对应实体的详细信息界面。用户可以通过控制面板来改变网络的布局，还可以放大缩小整个网络，以及下载网络图片。为了避免网络中节点数目过多，SymMap 在网

Summary of the herb: SMHB00264			
Chinese name	麻黄	**Pinyin name**	Ma Huang
Latin name	*Herba Ephedrae*	**English name**	Ephedra
Properties	Warm,Pungent,Slightly Bitter	**Meridians**	Lung,Bladder
Class in Chinese	辛温解表药	**Class in English**	Pungent-Warm Exterior-Releasing Medicinal
Use part	herbaceous twigs		
Function	To induce perspiration for dispelling cold, to relieve asthma, and to cause diuresis.		
External Links	TCM-ID:493 TCMID:2425 TCMSP:264		

图 3-4 检索结果页面的摘要面板

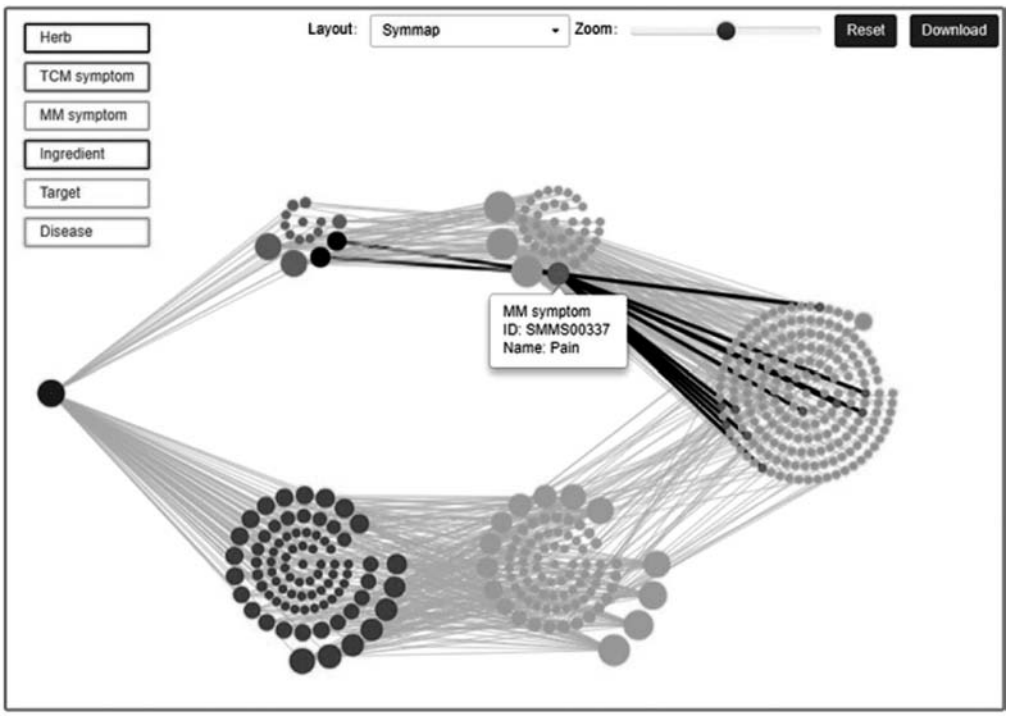


图 3-5 检索结果页面的网络面板

络面板仅展示错误发现率(False Discovery Rate,FDR)(Benjamini-Hochberg 方法)<0.05[8]的实体间的非直接关联。

5) 关联列表

关联列表展示网络可视化的信息,包括展示其检索项与其他五类实体之间的关联信息,如图 3-6 所示。首先,用户可以选择查看检索项与其他五类实体的其中一类的关联关系。第二,用户可以选择呈现不同严格程度的统计分析结果。第三,用户还可以按照 SymMap IDs、P-values、FDRs (BH) and FDRs (Bonferroni)等选项为结果排序。最后单击 Download 按钮可以下载调整后的关联列表。

3. 特性

SymMap 重点关注证候与中医药数据之间的关联。通过建立中医症状与西医症状以及西医疾病之

图 3-6　检索结果页面的关联列表：关联列表中的关联关系与图 3-5 中的网络可视化的内容相互对应

间的关系，进而搭建传统医学与现代医学之间的关联。此外 SymMap 还定量地描述了实体数据之间的关系，为中医药实体数据之间的关联研究提供了数据。

3.1.3　BATMAN-TCM：中药分子机制的生物信息学分析平台

BATMAN-TCM[7]（A Bioinformatics Analysis Tool of Molecular mechanism of Traditional Chinese Medicine）是一个中药作用机制的在线分析平台，用以揭示中药物质基础与人体生理过程之间的复杂相互作用。BATMAN-TCM 的主要功能包括：①中药成分的靶标预测；②靶标的功能分析；③成分-靶标-通路/疾病的相互作用网络可视化；④多个中药的比较分析。BATMAN-TCM 应用于预测芪参益气滴丸可能作用于肾素血管紧张素系统，后续通过相应的实验验证了芪参益气滴丸确实通过调控肾素血管紧张素系统发挥心肌保护的功能。BATMAN-TCM 致力于利用“多成分-多靶点-多通路”的整合策略来揭示中药的作用机制，通过该平台的预测为后续的实验验证提供有价值的线索，进而推动中药作用机制的研究。

1. 数据结构

BATMAN-TCM 支持三种类型的输入：①中药复方的拼音名，例如藿香正气散（huo xiang zheng qi san）；②中草药列表，拼音名、英文名或者是拉丁名均可（ren shen，Ginseng 或 Panax ginseng）；③化合物列表，要求输入 PubChem_CID 或 InChI 格式的化合物结构。上述的三种输入类型，包括复方、中草药和化合物，BATMAN-TCM 都会从后台的数据库中检索它们的组成化合物，用于后续的分析。

参数设置：Score_cutoff：默认为 20。针对每一种化合物，预测的候选靶标都会被赋予一个靶标预测的打分，得分范围为[0，1000]，仅打分大于 20 的潜在靶标（包括已知靶标）才会被纳入后续的功能分析中。

Adjusted P-value：默认为 0.05。统计学显著富集的功能条目的判断是基于该参数的。只有当 Adjusted P-value 小于用户所设置的数值时，该功能条目的富集才会被认为是有统计学意义的。

Adjusted P-value 指的是利用 Benjamini-Hochberg 多种检验校正后的 P-value[8]。

2. 功能介绍

1）功能一：中药成分靶标预测

针对用户输入的中药组成化合物，BATMAN-TCM 通过靶标预测后，会得到打分大于 Score_cutoff 的靶标列表，这些靶标被认为是符合筛选条件的潜在靶标（如图 3-7 所示）。后续都是基于这一步的潜在靶标结果分析。Score_cutoff 可以由用户在提交分析时设定，也可以在结果页面进行调整。

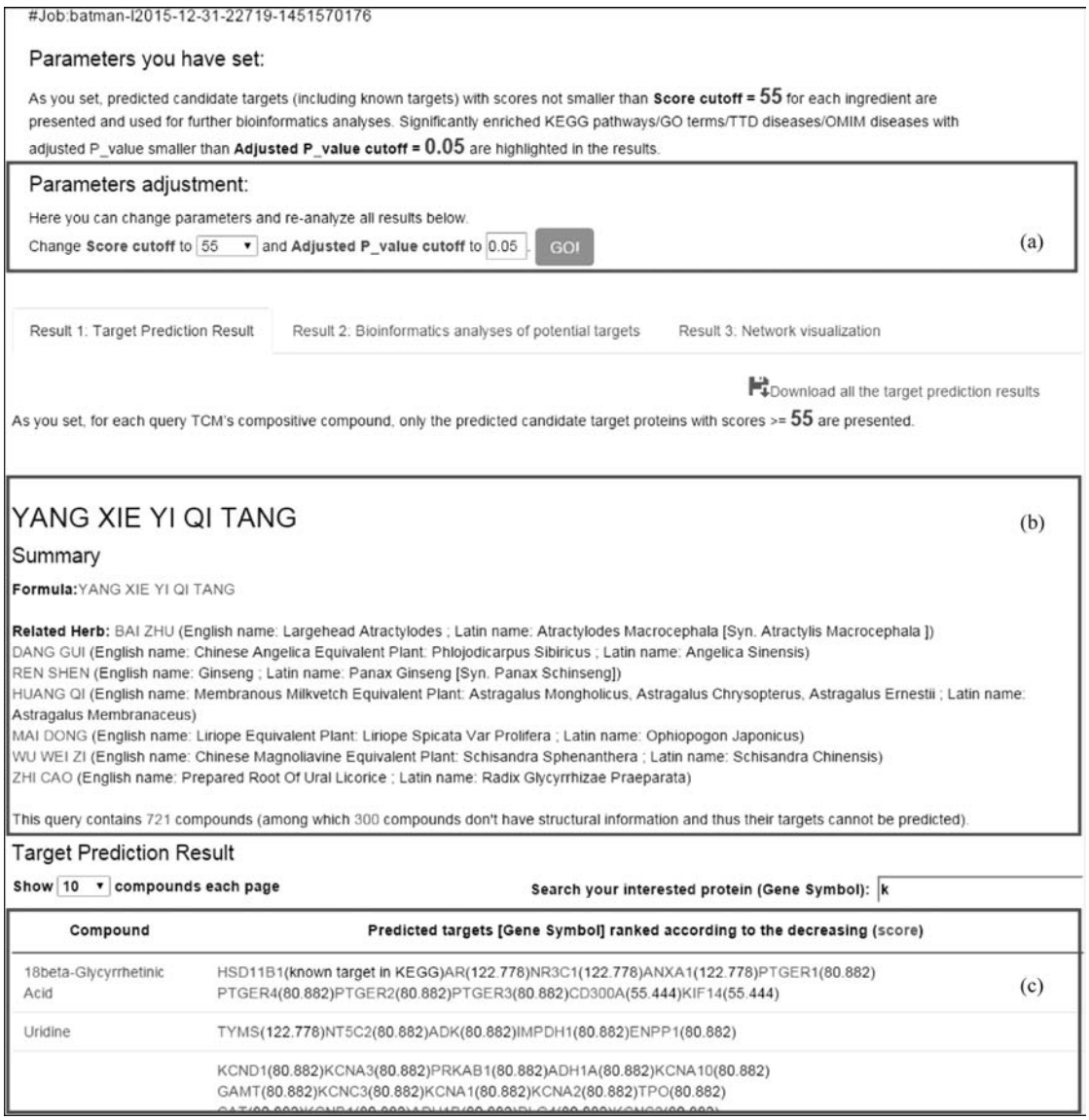


图 3-7 中药靶标预测结果页面

（a）在该页面参数 Score_cutoff 和 Adjusted P-value cutoff 都可以重新设定。一旦上述两个参数调整后，所有的分析结果都会随之更新。（b）用户输入概述，包括用户输入的复方名称、组成的中草药以及检索化合物列表。（c）靶标预测表格：在该表格中每一个成分都会列出其潜在作用靶标以及预测打分。此外 DrugBank，KEGG 或 TTD（Therapeutic Target Database）数据库中已收录的潜在靶标被标记为已知靶标

2）功能二：潜在靶标功能分析

该功能可以针对潜在靶标进行 KEGG 通路、GO 功能条目以及 OMIM/TTD 的疾病表型进行富集分析（如图 3-8 所示）。基于用户设置的 adjusted P-value 参数判断条目是否富集。在功能富集分析的结果表格中，富集的条目对应的 adjusted P-value 以及此条目包含的潜在靶标数量和列表均详细列出。针对 KEGG 通路富集的结果，还额外提供了潜在靶标在该通路的覆盖图。

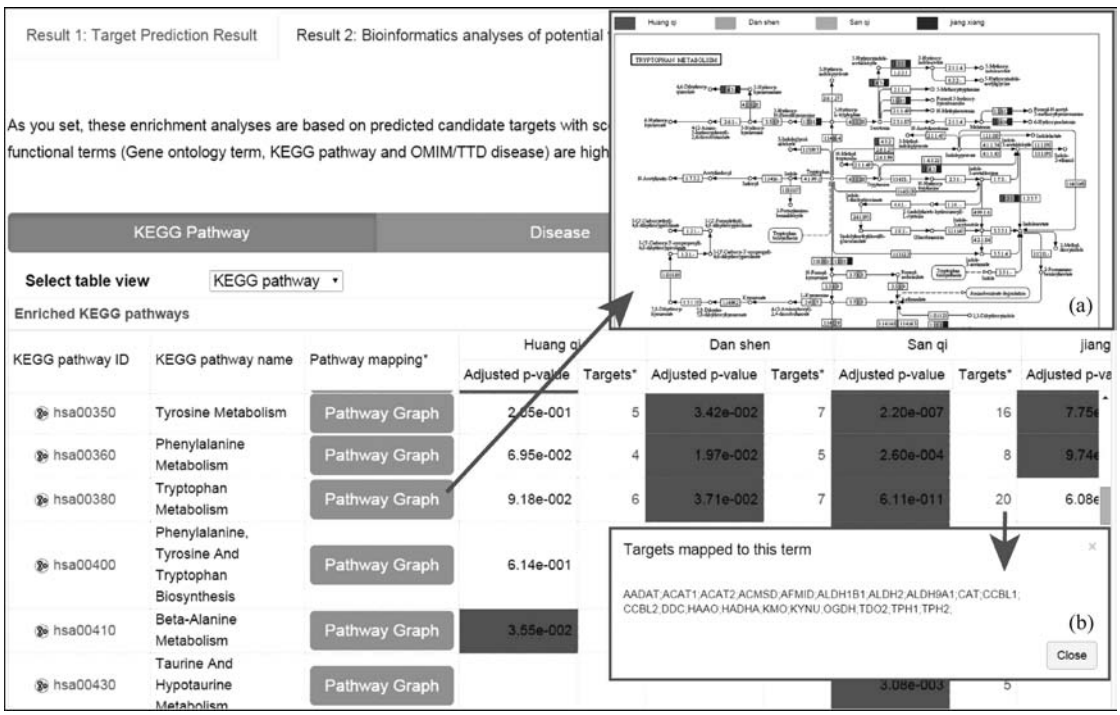


图 3-8　靶标富集分析结果页面

该页面呈现了潜在靶标显著富集的 KEGG 通路、GO 功能条目以及疾病条目。Adjusted P-value 小于卡值的功能条目被标记为红色。（a）在 KEGG 通路的结果页面单击“Pathway Graph”按钮，呈现的是潜在靶标在 KEGG 通路图上的覆盖情况。（b）单击“Target”字段下的数字可以呈现详细的该功能条目包括的潜在靶标列表

3）功能三：成分-靶标-通路/疾病的相互作用网络可视化

成分-靶标-通路/疾病的网络视图中，呈现了三种类型的关联，分别是用户输入的中药成分及其潜在靶标之间的关联、潜在靶标与生物学通路的关联、潜在靶标与富集的疾病条目的关联（如图 3-9 所示）。该可视化网络还可以通过修改潜在靶标的关联化合物数量来进行调整，目的在于聚焦重要的靶标及其相关功能之间的关联。

4）功能四：比较分析

用户可以同时提交多个任务进行分析，BATMAN-TCM 将从靶标、功能、网络等多方面提供计算结果的比较。在 BATMAN-TCM 中，每提交一个任务被定义为一个簇（cluster）。在靶标预测结果页面会提供不同 cluster 之间的靶标比较维恩图。而在功能富集分析结果页面会提供不同 cluster 在同一功能条目上的富集以及覆盖情况。

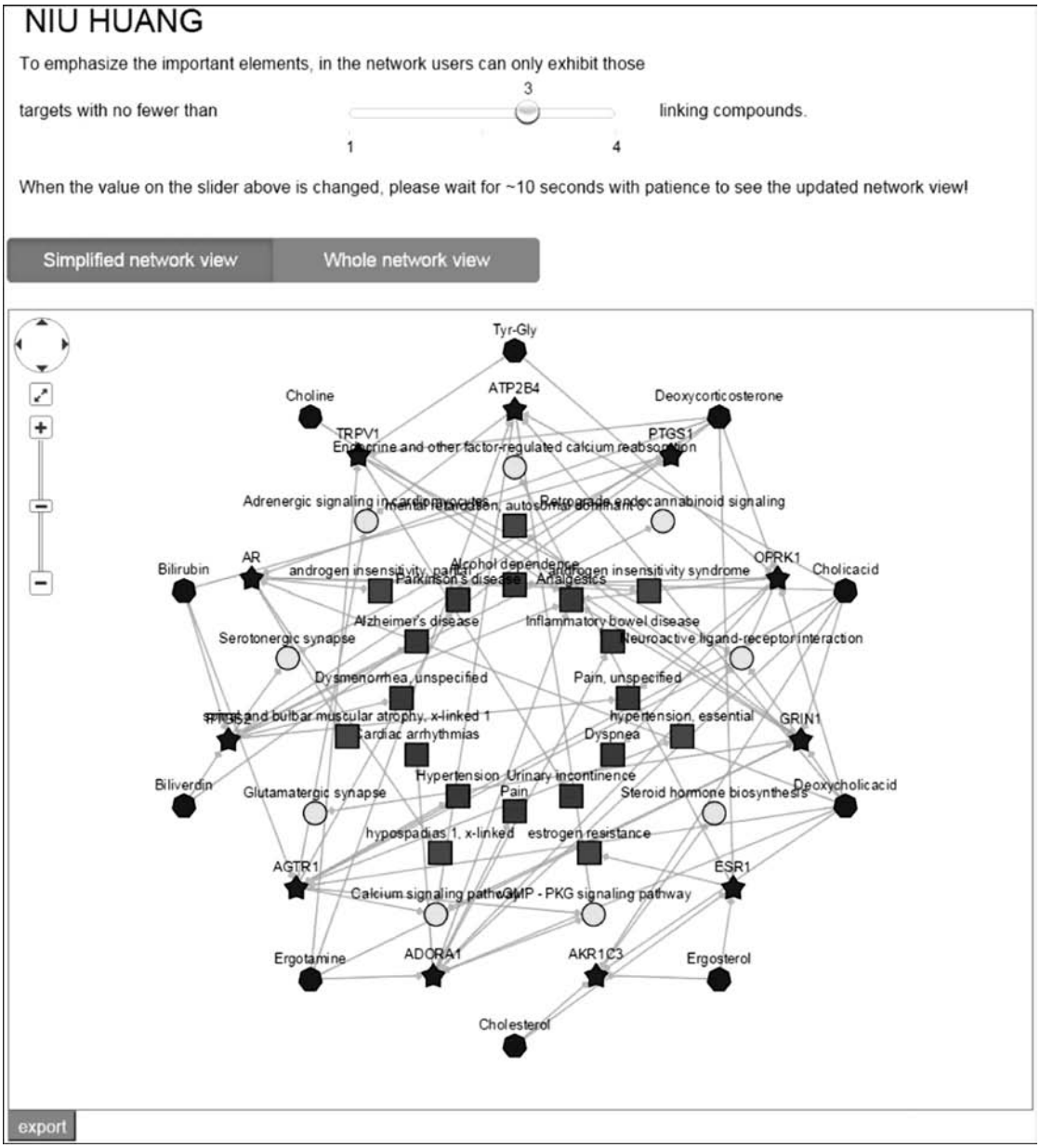


图 3-9 成分-靶标-通路/疾病关联网络

在该网络中，中药成分、潜在靶标、通路以及疾病为四种不同类型的节点，用不同的颜色和形状加以区分。此外还包括三种类型的关联，分别是用户输入的中药成分及其潜在靶标之间的关联、潜在靶标与生物学通路的关联、潜在靶标与富集的疾病条目的关联

5）功能五：通过功能检索中药

通过单击首页的“Function2TCM”按钮，用户可以寻找特定通路、疾病或者 GO 条目所关联的中药复方及中草药列表。

3. 特性

BATMAN-TCM 是由军事医学科学院贺福初院士团队开发的一个中药作用机制在线分析平台。该平台以数据分析见长，可以进行中药的靶标预测和功能分析。此外还提供了不同药物之间的比较分析，可以用于中药君臣佐使不同角色药物之间的比较以及通过通路检索中药的功能。

3.1.4　TCMID：用于中药分子机制分析的中医药整合数据库

TCMID[2,3]（Traditional Chinese Medicine Integrative Database）记录了从不同资源中通过文本挖掘方法收集到的中医相关信息。TCMID 由处方、药材、成分、靶点、药物和疾病六个数据字段组成，主要目标是通过疾病基因/蛋白质在中草药成分和疾病之间建立联系，这也可能是潜在的药物靶点。该平台基于网络的软件显示了一个网络，用于展示中草药和它们治疗的疾病、活性成分和它们的靶标之间的综合关系，这将促进联合治疗的研究，并在分子水平上理解中医的潜在机制。其建立的主要网络有：①中草药-疾病网络；②中草药成分-靶点相互作用网络；③中草药成分-靶点-疾病-药物网络。

1. 数据结构

六个数据字段在数据库系统的内在关系如图 3-10 所示。处方由草本植物组成，草本药物含有各种成分（化合物），一种成分（或药物）可以与它的靶标（蛋白质）相互作用，疾病可能是由基因/蛋白质功能引起的。

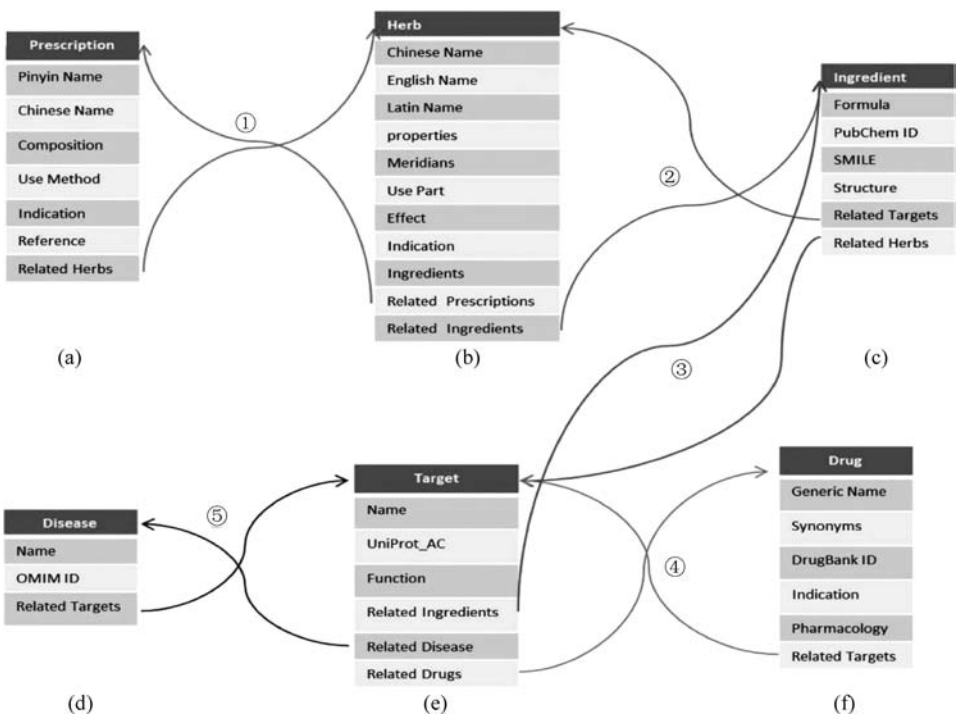


图 3-10　六个数据字段在数据库系统与内在关系

数据库结构：A～E：处方、药材、成分、疾病、靶点、药物六个数据字段

1～5 代表意义不同。①处方由中草药组成。②中草药含有成分。③成分与靶标相互作用。④靶标确定药物。⑤靶标引起疾病。

由于来自六个不同字段的信息和数据是连接的，用户可以使用任何数据字段查询数据库，并按照链接检索相关信息。

例如，用户可以选择 Herb 作为入口点，使用英文 Herb 的名称进行查询。结果页面将显示查询中草药的信息，并显示其与处方和中草药成分的链接，它们通过这些超链接与目标交互。

2. 功能介绍

1）功能一：中草药-疾病网络

在传统中医药中，中草药或方剂是在特定的“pattern”（中文称为“证”）下组方的，是一个特定的功能状态的描述，而现代医学的药物是为治疗某些疾病而设计的。因此，将中草药或方剂与其治疗的疾病联系起来是有必要的。由于某个致病基因/蛋白可能是某一草本植物成分的靶点，TCMID 基于这一观点构建了这样一个中草药-疾病网络（如图 3-11 所示）。

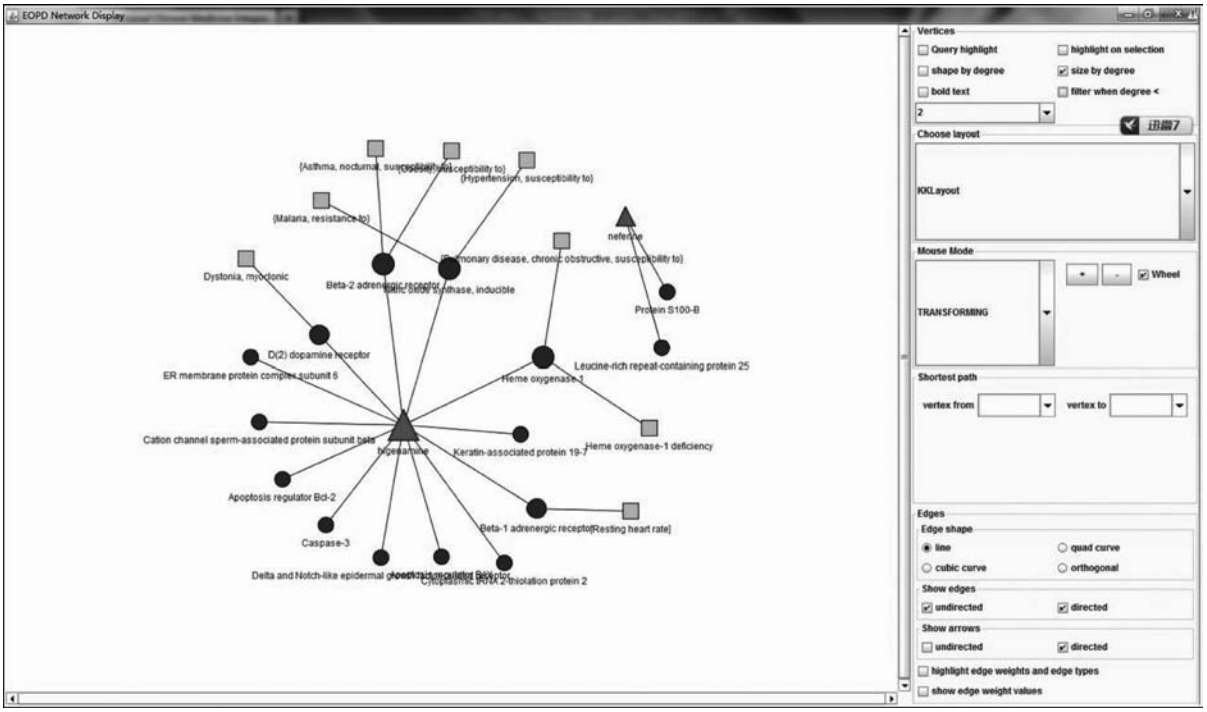


图 3-11 中草药-疾病网络

三角形：中草药成分，圆圈：中草药成分的靶标，方块：与疾病相关的靶点

节点大小取决于其等级

2）功能二：中草药成分-靶标相互作用网络

该网络是为了探索成分之间的相互作用，促进联合治疗的研究。在这个网络中，如果这些成分能够作用于实验数据或计算方法所支持的蛋白质，那么蛋白质和成分就会相互关联。因此，如果两种成分都能基于网络，与同一蛋白或存在相互作用的不同蛋白相互作用，则用户可以推断出两种成分之间潜在的

协同/拮抗作用(如图 3-12 所示)。但 TCMID 在成分与靶标的关联中存在一点缺陷,只能通过输入靶标查到对应的成分,而不能查到中药或者成分的作用靶标。

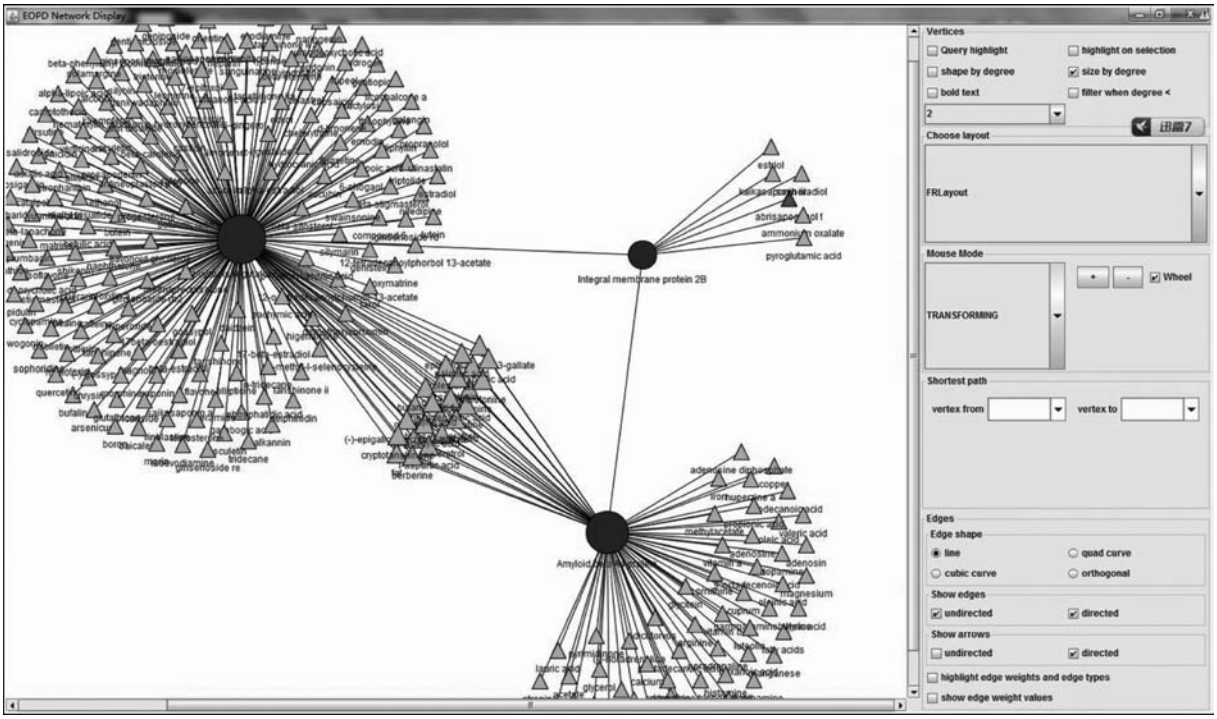


图 3-12　中草药成分-蛋白质相互作用网络

红色三角形,用于查询的成分,黄色三角形:中草药成分,蓝色圆圈:中草药成分的靶标

节点大小取决于其等级

3) 功能三:中草药成分-靶点-疾病-药物网络

为了探索中草药成分的潜在作用机制,该网络将这些成分与它们的潜在靶点、相关疾病和相关药物联系起来。此外,还建立了一个工具来显示一个网络中的关系,它为用户提供了一个直观的视图来推断疾病的治疗机制,并通过它们之间的联系来识别潜在的目标成分。如果一种中草药成分能与疾病有关的蛋白靶点相互作用,就表明该成分具有治疗疾病的潜在机制。此外,如果中草药成分具有与药物相同的靶点,就意味着该成分具有潜在的药理作用(如图 3-13 所示)。

3. 特性

TCMID 的一个优点就是包含的中药和成分的种类较为全面,其中收集了 8159 味中药和 43413 种化学成分。TCMID 数据库的当前版本虽然有“target”选项,但是只能由靶标查询化合物,不能由中草药或者化合物查询靶标,也不能展示出中草药-靶标-疾病网络。TCMID 系统为中医药现代化的分子水平机制研究提供了新的思路。随着中医对疾病的治疗更加全面,有必要采用系统的方法来探索中医的潜在机制和治疗效果。因此,该系统尝试通过现代西医和中医的共同之处——中草药成分/化合物与它们的目标相结合,使两千多年的临床实践积累的知识与现代经验或计算方法相结合。这一综合信息不仅有利于中医药,也将促进网络药理学的发展。此外,随着系统生物学的发展,越来越多的“组学”方法,如蛋白质组学和代谢组学,逐渐被中医研究采用。因此,收集这类信息必将有助于促进中医药的系统

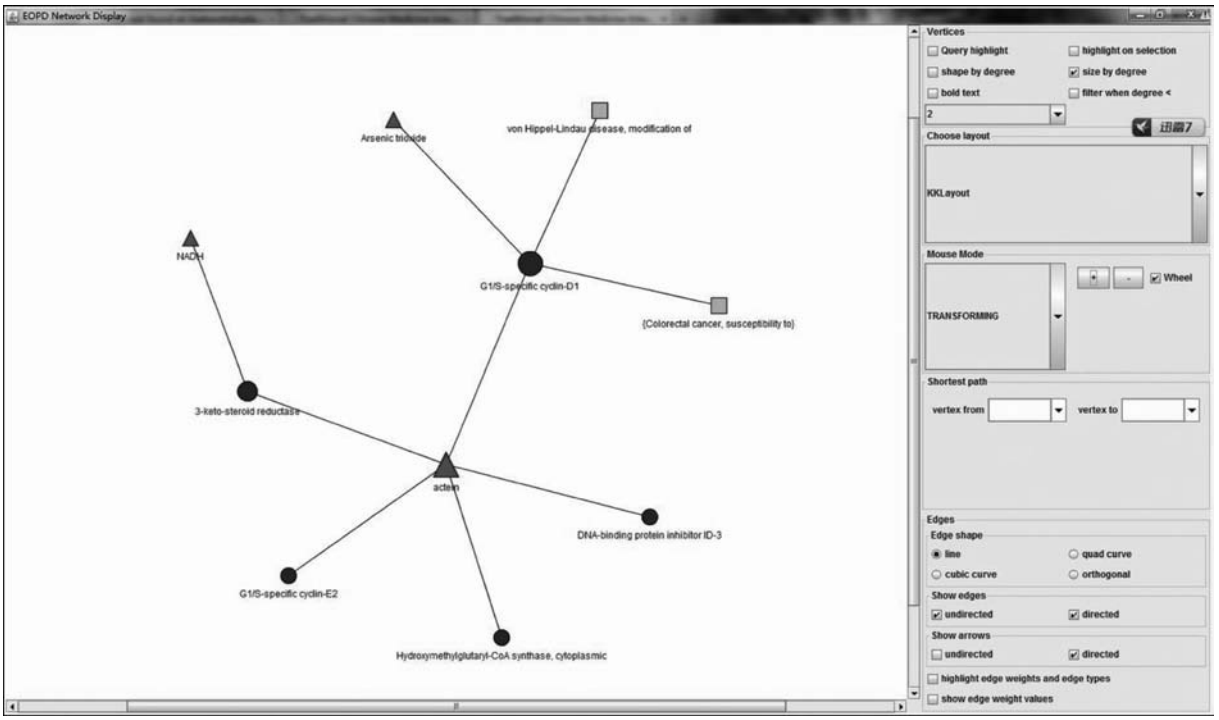


图 3-13 中草药成分-靶点-疾病-药物网络

红色三角形：草本成分，绿色三角形：药物，黄色正方形：疾病，蓝色圆圈：中草药成分的目标

节点大小取决于其等级

研究。

3.1.5 其他中医药数据库

TCM-ID[9]作为一个信息平台，提供有关中医药各方面的信息，包括处方、组成每个处方的中草药、中草药成分、活性化合物的分子结构和功能属性、每个处方的临床适应症和应用、中药成分的疗效和毒性效应及相关文献。目前，TCM-ID 中包含了 1588 个处方、1313 种中草药、5669 种中药成分以及 3725 种药物成分的三维结构。TCM-ID 中数据的价值在于一定程度上可以解决诸如中药机理研究等问题[10]。为了解决这些问题，两项独立的研究中使用了 TCM-ID 的一些数据。首先，利用特定中草药成分的三维结构，经由电脑模拟计算预测它们的分子靶点。对已确定的靶点进行进一步检测，以确定这些成分的已知治疗效果是否可以通过干扰这些靶点的预期效果来解释。其次，利用已知的中药处方开发了一个人工智能（AI）系统来验证新的中药复方制剂。所开发的人工智能系统使用了一些尚未包含在 TCM-ID 中的新出版的中药处方进行测试。

TCMSP[6]是以中药系统药理学为框架建立的中药系统药理学数据库与分析平台，包含每种活性化合物的药物靶点和作用疾病，可以自动建立化合物靶点和靶点疾病网络，让用户查看和分析药物作用机制。TCMSP 旨在推动中草药的发展，促进现代医学和传统医学的结合，促进药物的发现和开发。TCMSP 的特点在于其包含了大量的中草药成分，以及具有计算预测药物靶点网络和药物疾病网络的能力，有助于揭示中药及中药配方的潜在作用机制，发现药物和药物组合。

TCMGeneDIT[5]是一个提供中药、基因、疾病、中医功效和中药成分关联关系的数据库，这些关联关系是来自中国台湾大学的研究人员从海量的生物医学文献中挖掘到的。中药、基因和疾病之间的关系可以通过中间对象的传递来进行考察。整合蛋白和蛋白相互作用与生物学通路的信息也被用来考察与中药作用有关的基因调控关系。TCMGeneDIT 通过基因调控关系并推导协同作用和拮抗作用的贡献来帮助人们理解中药可能的作用机制。

3.2　网络药理学常用生物相关数据库

近年来，各类生物学数据库的构建为网络药理学研究提供了可靠有力的数据支撑。网络药理学研究常用的生物学数据库包括疾病表型与基因型关联数据库（OMIM，HPO，DisGeNET）、药物靶标信息数据库（TTD，PDB，KEGG）和生物分子相互作用数据库（BioGRID，DIP，IntAct，MINT，STRING）等。基于这些包含临床和基础研究结果的生物学数据库，网络药理学研究即可构建“疾病表型-基因-靶点-药物”相互作用网络，并在此基础上考察药物对疾病网络的干预特点和作用机制。

3.2.1　OMIM：人类孟德尔遗传病在线数据库

1. 数据库内容及其在网络药理学研究中的应用

OMIM（Online Mendelian Inheritance in Man）数据库是一个有关人类基因与遗传性状的综合性权威数据库，此数据库重点关注疾病表型与基因型之间的联系[29]，其收录了所有孟德尔遗传性疾病和超过 15 000 种人类基因的相关信息，包括所有已知的遗传病、遗传决定的性状及其基因，除了简略描述各种疾病的临床特征、诊断、鉴别诊断、治疗与预防，还提供已知的致病基因的连锁关系、染色体定位、组成结构和功能、动物模型等信息，并附有经人工核查的相关文献证据[11]。OMIM 制定的各种遗传病、性状、基因的编号（数据分类及条目详如表 3-1 所示），简称 OMIM 号。有关疾病的报道必须冠以 OMIM 号，以明确所讨论的是哪一种遗传病[12]。OMIM 数据库为网络药理学研究提供详细、实时更新及可免费下载的疾病相关基因数据，为构建和挖掘疾病相关基因与药物靶标基因的互作关联性提供可靠的数据支撑。

表 3-1　OMIM 整体数据情况

MIM 标识	常染色体遗传	X 连锁遗传	Y 连锁遗传	线粒体遗传	总计
明确的基因座	15 281	733	49	37	16 100
已知表型的基因座	44	0	0	0	44
通常有表型的描述	5195	336	5	33	5569
分子机制未知的孟德尔遗传性状	1438	119	4	0	1561
主要表型是否为孟德尔表型尚未确定	1644	105	3	0	1752
总计	23 602	1293	61	70	25 026

OMIM 不仅收录了以孟德尔方式遗传的所有单基因病的相关资料，还收录了染色体病、多基因病、线粒体病方面的资料，所涵盖的病种异常丰富。具体到每一条目也即是每一种疾病，OMIM 都提供了从基础到临床的全方位的信息。具体信息包括基本描述（Description）、基因定位（Mapping）、分子遗传

学(Molecular Genetics)、遗传方式(Inheritance)、基因定位(Mapping)和群体遗传学(Population Genetics)等。并且每一方面的描述都提供了相应参考文献的链接,可供进一步查阅。

2. 功能介绍

遗传病种类繁多,但就特定病种来说又较为罕见,临床医师和遗传学专家难以对每种遗传学疾病都了然于胸。OMIM 提供了大量孟德尔遗传病的临床特征(Clinical Feature)、诊断(Diagnosis)、临床治疗方案(Clinical Management)和基因治疗(Gene Therapy)等方面的信息。简明扼要的临床纲要(Clinical Synopsis)具有很强的实用性,与其他数据库的链接可以获得更多相关信息,如 GeneTests,可以提供多种遗传疾病的诊断试验信息。OMIM 对临床和遗传咨询工作者来说,犹如一个功能强大的专家系统。同时 OMIM 可以提供关于特定疾病(包括多基因遗传病)临床表型和致病基因多方面的信息,包括基因定位、分子机制、病理、动物模型、遗传方式等。并且每一方面的描述都提供了相应参考文献的链接,使研究人员可以快速而全面地把握某种疾病的主要信息和最新进展。

3. 特性

OMIM 是一个关于人类基因和遗传疾病的综合性数据库,其收录了所有的孟德尔遗传性疾病和人类基因信息,除了简略描述各种疾病的临床特征、诊断、治疗与预防外,还提供了已知有关致病基因的连锁关系、染色体定位、功能、动物模型等资料并附有经缜密筛选的相关参考文献。及时性、权威性、全面性和实用性是 OMIM 的特点。但是 OMIM 数据库模式(Database Schema)和数据模型(Data Model)不透明,所以无法利用 SQL 自行编写查询语句进行数据库的知识发现。对于复杂性疾病,例如哮喘,由于所需分析的数据类型异常复杂,OMM 目前提供的解决方案无法满足日益增长的研究需要。而且 OMIM 包含的内容没有座位专一数据库(Locus-Specific Database)那样丰富:突变数据没有完全收集,缺乏引物设计的信息、基因表达谱等。数据注解仅限于遗传学方面。

3.2.2 HPO:人类表型本体数据库

1. 数据库内容及其在网络药理学研究中的应用

HPO(Human Phenotype Ontology)数据库由柏林 Charité 大学医院 Peter N. Robinson 与 Sebastian Köhler 于 2007 年创立,提供医学相关表型、疾病表型注释以及基于表型的本体信息。HPO 术语覆盖解剖学、细胞类型、生物功能、胚胎学、病理学等众多领域的 13 000 多个术语和 156 000 多个遗传病注释。大多数本体信息以有向非循环图(Directed Acyclic Graphs,DAG)形式构成。例如,术语"跖骨发育不全/发育不良"是指涉及儿童脚骨的发育不全/发育不良和异常跖骨形态。在 DAG 中编码多个父类术语增加了本体信息的灵活性和描述性。术语的子父关系是可传递的,这意味着注释继承到根的所有路径。例如,左心室形态异常-心室形态异常。目前,HPO 被广泛应用于计算深度表型和精准医学,将临床数据整合到转化研究中,已被多家国际罕见疾病组织、注册管理机构、临床实验室、生物医学资源和临床软件工具等不同群体作为判断表型异常的标准[13~15]。在网络药理学研究中,HPO 可提供具体的疾病症状表型描述及其相关基因集,可满足用户按照目标疾病的病理环节收集相关基因信息,从而探索药物对目标疾病进展过程中某个病理环节的网络调控机制。HPO 涵盖的表型分类如表 3-2 所示。

表 3-2　HPO 涵盖的表型分类

表 型 类 型	词条举例说明
形态学异常(Morphological Abnormality)	Arachnodactyly(HP：0001166)
器官功能异常【Abnormal Process (organ)】	Epistaxis(HP：0000421)
细胞功能异常【Abnormal Process (cellular)】	Abnormality of Krebs Cycle Metabolism(HP：0000816)
实验室指标异常(Abnormal Laboratory Finding)	Glycosuria(HP：0003076)
电生理指标异常(Electrophysiological Abnormality)	Hypsarrhythmia(HP：0002521)
医学影像学指标异常(Abnormality by Medical Imaging)	Choroid Plexus Cyst(HP：0002190)
行为学异常(Behavioral Abnormality)	Self-Mutilation(HP：0000742)

2. 数据结构

HPO 的每个术语都描述一种临床表型。这些术语可能是一般术语，如异常耳朵形态，也可能是专业术语，如脉络膜视网膜萎缩。每个术语也分配给五个子本体信息，即表型异常(Phenotypic Abnormality)、遗传方式(Mode of Inheritance)、临床干预(Clinical Modifier)、临床病程(Clinical Course)和表型出现频率(Frequency)。这些术语均具有唯一的标识，即 HPO 标签，如“HP：0001140”表示“球外胚层”。该数据库对于大多数的表型具有具体的定义和描述，且提供证据来源。例如，球上皮样瘤是一种良性肿瘤，通常发现于角膜和巩膜的交界处(角膜缘球上皮样瘤)。

3. 功能介绍

HPO 常被用于临床诊断、基于表型的基因组诊断、生物信息学数据挖掘等多种不同的工具和算法[16]。常用工具如下。①临床诊断工具：HPO 为实现模糊、特异性加权表型匹配的算法提供了计算基础，以支持微分诊断。用户可以单击 HPO 术语列表中代表表型异常(体征、症状、实验室测量等)查询。②外显子组/基因组诊断和研究工具：HPO 基于算法基础开发了一系列 Java 的工具，旨在实现孟德尔病变异表型驱动优化。这些工具可以导出与表型异常相关的外显子组或基因组中提取的 VCF 文件与 HPO 术语列表。③拷贝数变异诊断工具：微阵列比较基因组杂交技术和相关检测通常被用作发育迟缓和先天性畸形等适应症的筛查试验。这些检测可以检测拷贝数变异(删除和复制)。在所有个体中都可以发现大量的拷贝数变异，因此，很难确定拷贝数变异是否与疾病相关。基于 HPO 可以分析基因拷贝数变异是否与患者身上观察到的表型异常相关，进而确定拷贝数变异与疾病的相关性。④临床表型工具：HPO 侧重于准确的临床表型分型，以促进疾病分类和候选标志基因的发现。

4. 特性与不足

HPO 为研究人员和临床医生提供了定义明确、较为全面且可互操作的疾病表型数据资源，并在临床和研究环境中被用作疾病表型分析的基础工具，集成了不同学科和数据库的复杂表型信息。最初 HPO 术语的重点是罕见的疾病，主要是孟德尔遗传病；虽然 HPO 术语现在也可用于常见疾病，目前其资源也已覆盖精密医学、癌症和以非孟德尔遗传病，但仍需进一步扩大其覆盖范围。

3.2.3　DisGeNET：疾病基因关联数据库

DisGeNET 数据库是将疾病和基因的关联信息及相关药物信息相整合的开源数据库，其证据来源

于其他数据库和文献。当前版本的 DisGeNET(v6.0)包含 17 549 个基因与 24 166 个疾病、异常、性状和临床或异常的人类表型，共 628 685 个基因-疾病关联(GDAs)；同时还有 210 498 个变异-疾病关联(VDAs)，由 117 337 个变异体与 10 358 种疾病、性状和表型组成。GDAs 的信息来源主要由以下四部分组成。①CURATED：由 UniProt，PsyGeNET，Orphanet，CGI，CTD (Human Data)，ClinGen 和 Genomics England PanelApp 专业数据库提供 GDAs 信息；②ANIMAL MODELS：RGD，MGD 和 CTD (mouse and rat data)，这些数据包括动物模型(目前为大鼠和小鼠)疾病信息的资源提供的 GDAs，并且使用同源性分析来映射与人类基因的关联；③INFERRED：此部分数据指从 Human Phenotype Ontology(HPO)和 VDAs 推断出来的 GDAs，其数据库来源包括 HPO，CLINVAR，GWAS Catalog 和 GWAS DB；④LITERATURE：包括 LHGDN 和 BeFree 数据库。VDAs 信息来源主要包括：①CURATED：包括 UniProt，ClinVar，GWAS Catalog 以及 GWAS db 数据库；②INFERRED：SETH 工具。作为一个多功能信息平台，DisGeNET 数据库已被广泛用于人类疾病及其并发症的分子基础研究、疾病基因特征挖掘、药物治疗作用和药物不良反应的生物学基础研究，以及针对计算预测所得疾病基因的验证和文本挖掘方法性能的评估[21,22]。

1. 数据结构

为了集成基因疾病关联数据，DisGeNET 数据库开发了关联类型本体。如果基因/蛋白质与疾病之间存在关系，那么在原始源数据库中发现的所有关联类型都由父类基因-疾病关联(gene-disease association)类正式构造，并表示为本体类。它是已经集成到 Semantic Science Integrated Ontology (SIO)中的本体论语言(Web Ontology Language，OWL)本体，为丰富的对象、过程及其属性描述提供了必要的类型和关系。

2. 功能介绍

DisGeNET 数据库中，大多数 GDAs 通过使用 BeFree 文本挖掘文献进行识别，并整合各种权威来源的人类遗传学数据库。每个 GDA 都使用其支持证据进行明确注释，这使 DisGeNET 成为基于证据的知识发现的参考资源[23,24]。DisGeNET 包含与疾病相关的基因汇编，且来自不同的开源数据库。通过 DisGeNET 可以获得基因变异相关的疾病信息；疾病与基因的关联信息；特定基因与疾病之间的关联类型；针对某种特定疾病，最新发现的相关基因和变异信息。

3. 特性

DisGeNET 数据库的亮点是数据集成、标准化和对证据来源的跟踪查询功能。通过基因和疾病词汇表映射以及使用 DisGeNET 关联类型本体来执行整合。此外，GDAs 根据其类型和证据水平进行组织，如 CURATED，PREDICTED 和 LENTERATURE，并且还根据支持证据对其进行评分，以确定优先次序并减少其探索不足。DisGeNET 的目标是整合所有疾病的遗传基础信息，成为一个参考知识库，以填补基因型和表型差异，目前 DisGeNET 平台被用来研究生物医学问题。

3.2.4 MalaCards：疾病信息数据

MalaCards(Mala Cards Human Disease Database)数据库是魏茨曼科学研究所和美国犹他谷大学联合开发的综合数据库，其整合了各大数据库网站的人类疾病及其注释数据[33]。该数据库是从 68 个

数据源中挖掘出的综合疾病概要，包含全球 6 个类别的 20000 个疾病条目。每个疾病都包含了 14 个小部分注释，包括总结、症状、解剖背景、药物，基因测试、突变信息以及和该疾病相关的一些文献等。该数据库能够将来自互补来源的信息相互整合结合其精准的搜索功能、关系数据库基础设施和方便的数据转储功能，使其能够处理丰富的疾病注释资源，并有助于系统分析和基因组序列解释等功能。

1. 数据库介绍及使用

1）疾病查询

该数据库整合了 75 个数据库的信息，对相应疾病的搜索只需输入疾病名称即可。搜索界面如图 3-14 所示。

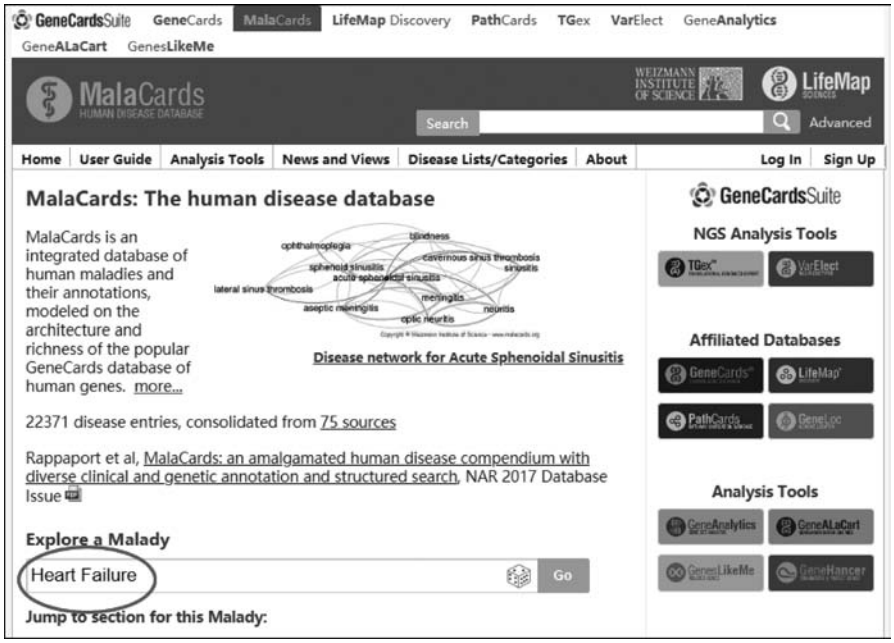


图 3-14　疾病搜索界面

以某一疾病为例，MalaCards 把关于该疾病的搜索结果分成了 14 个小部分（如图 3-15 所示），可直接单击“Jump to Section”跳转到该部分，同时 MalaCards 的 Summaries 部分可以查看多个数据库对于这个疾病的总结（如图 3-16 所示）。

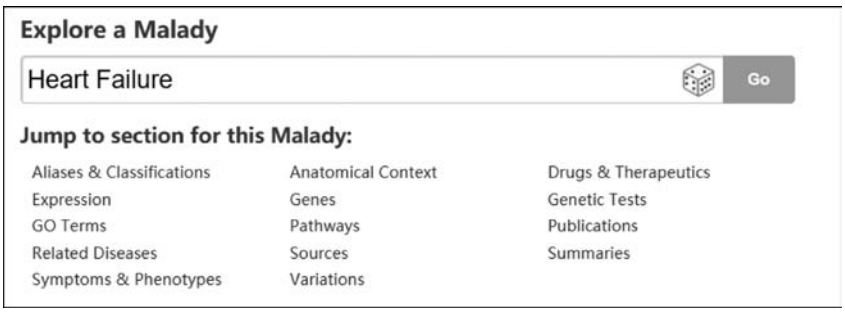


图 3-15　疾病的 14 个小部分

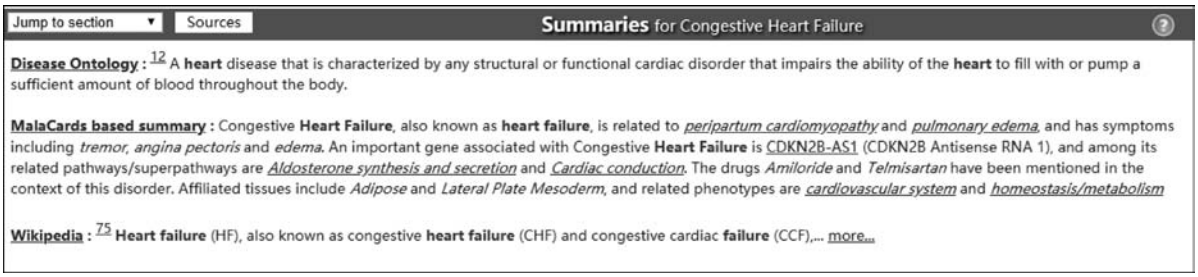


图 3-16 疾病小结

2）网络分析

MalaCards 构建了表型-疾病网络，可以看到相互联系疾病之间的关系（如图 3-17 所示）。

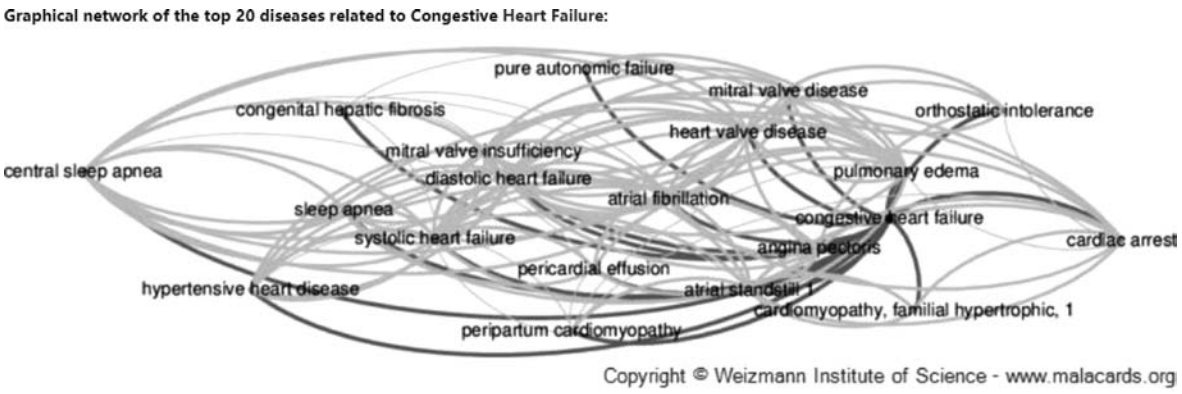


图 3-17 表型-疾病网络

3）疾病相关知识拓展

（1）目前研究治疗疾病的一些药物以及治疗方法如图 3-18 所示。

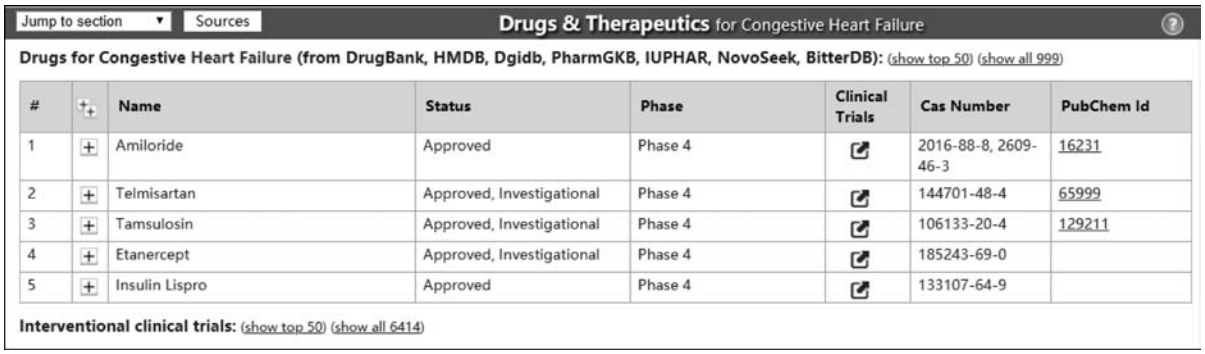


图 3-18 已知的疾病治疗方法以及药物

（2）与心衰相关的文章如图 3-19 所示。

（3）MalaCards 可获得与疾病相关的关键基因，如图 3-20 所示。

（4）与疾病相关的各大网站入口如图 3-21 所示。

Jump to section ▾　Sources　**Publications** for Congestive Heart Failure

Articles related to Congestive Heart Failure: (show top 50) (show all 54425)

#	Title	Authors	PMID	Year
1	Provider adherence to clinical guidelines related to lipid-lowering medications. [9 38]	Cohen SM...Kataoka-Yahiro M	20180482	2010
2	Structure of human G protein-coupled receptor kinase 2 in complex with the kinase inhibitor balanol. [9 38]	Tesmer JJ...Huber J	20128603	2010
3	Effects of eythropoietin administration on mitral regurgitation and left ventricular remodeling in **heart failure** patients. [9 38]	Cosyns B...Lancellotti P	18760492	2010
4	Potential of endothelin-1 and vasopressin antagonists for the treatment of congestive **heart failure**. [9 38]	Rehsia NS...Dhalla NS	19763821	2010
5	Alterations in plasma semicarbazide-sensitive amine oxidase activity in hypertensive **heart** disease with left ventricular systolic dysfunction. [9 38]	Marinho C...Bicho M	20391898	2010
6	Crystal structure of the sodium-potassium pump (Na+,K+-ATPase) with bound potassium and ouabain. [9 38]	Ogawa H...Toyoshima C	19666591	2009
7	Acute hemodynamic effects of intravenous sildenafil citrate in congestive **heart failure**: comparison of phosphodiesterase type-3 and -5 inhibition. [9 38]	Botha P...Macgowan GA	19560695	2009
8	The clinical pharmacology of eplerenone. [9 38]	Muldowney JA...Benge CD	19379127	2009
9	Coupling of a vented column with splitless nanoRPLC-ESI-MS for the improved separation and detection of brain natriuretic peptide-32 and its proteolytic peptides. [9 38]	Andrews GL...Muddiman DC	19269262	2009
10	A pilot study on the role of autoantibody targeting the beta1-adrenergic receptor in the response to beta-blocker	Nagatomo Y...Ogawa	19327624	2009

图 3-19　疾病相关文章

Jump to section ▾　Sources　**Genes** for Congestive Heart Failure

Genes related to Congestive Heart Failure (0 elite genes): (show all 29)

★ - Elite gene　Ⓒ - Cancer Census gene in COSMIC

#	Symbol	Description	Category	Score	Evidence	PubmedIds
1	CDKN2B-AS1	CDKN2B Antisense RNA 1	RNA Gene	150	Experimental evidence: Expression [39]	27317124
2	HOTAIR	HOX Transcript Antisense RNA	RNA Gene	150	Experimental evidence: Expression [39]	27317124
3	TUSC7	Tumor Suppressor Candidate 7	RNA Gene	150	Experimental evidence: Expression [39]	27317124
4	NPPA	Natriuretic Peptide A	Protein Coding	45.8	DISEASES inferred [15] Novoseek inferred [55] GeneCards inferred via (show sections)	1473654 8376700 9398103 (more)
5	VCL	Vinculin	Protein Coding	40.34	DISEASES inferred [15] GeneCards inferred via (show sections)	
6	ADRB1	Adrenoceptor Beta 1	Protein Coding	39.43	DISEASES inferred [15] Novoseek inferred [55] GeneCards inferred via (show sections)	12422153 1600963 12463096 (more)
7	ACE	Angiotensin I Converting Enzyme	Protein Coding	39.31	DISEASES inferred [15] Novoseek inferred [55]	8682064 18458262 8863101 (more)

图 3-20　疾病关键基因

2. 特点

（1）数据库资源丰富，来源于 75 个数据库。

（2）疾病相关拓展知识丰富，可以从包括文献等多个方面对该疾病有进一步的了解。

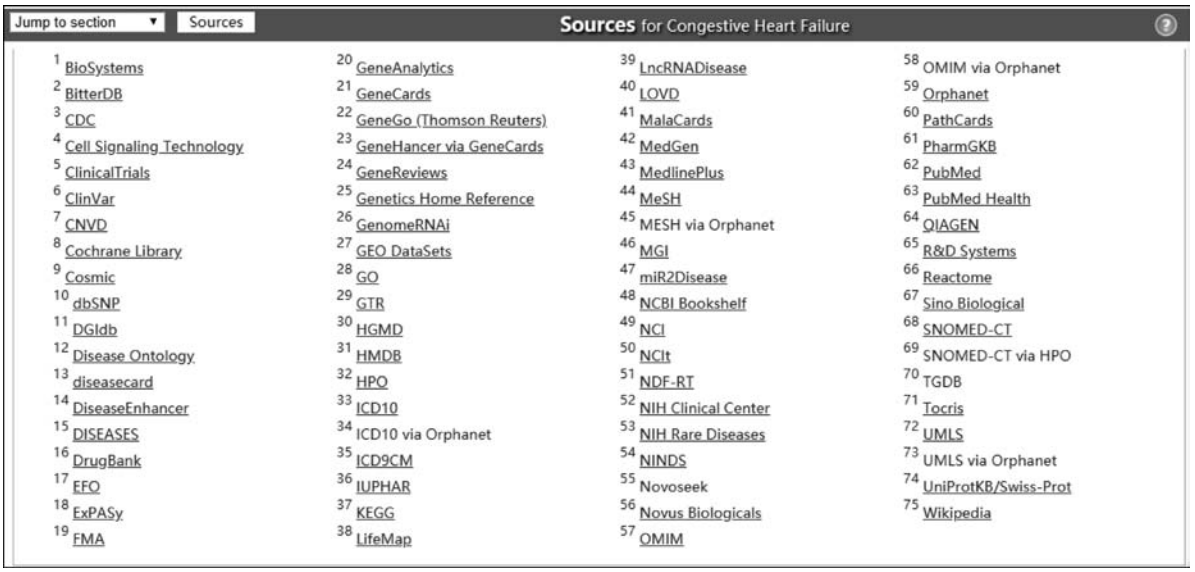


图 3-21 疾病相关网站入口

3.3 网络药理学常用靶标相关数据库

3.3.1 TTD：治疗靶标数据库

TTD(Therapeutic Target Database)数据库由新加坡国立大学生物信息与药物设计(Bioinformatics & Drug Design,BIDD) 研究团队创建而成，该数据库最近一次更新的时间是 2017 年 9 月 15 日。根据 2018 年 TTD 更新统计显示：数据库涵盖的药物靶标共 3101 个，其中成功验证的靶标 445 个，用于临床的治疗靶标 1121 个，处于研究阶段的靶标 1535 个；数据库共收录 34 019 种药物，被临床批准使用的药物共计 8103 种，正在进行研究的药物 18 923 种，多靶点制剂 26 459 种，退出市场的药物 158 种，临床停止使用的药物 2349 种，前期临床实验药物 417 种，在未指定研究阶段终止的药物 1929 种，有效的小分子药物 21 936 种，被批准的具有有效结构的药物 2326 种，可用于临床试验结构的药物 4258 种，对现有结构研究的药物 15 352 种；此外，更新后的数据库还增加了双特异性抗体 21 种，干细胞药物 10 种。该数据库提供已知或探索中治疗蛋白靶点和核酸靶点相关信息，此类靶点针对的疾病、通路信息和对应的药物配体分子等。TTD 数据库还提供相关数据库的链接，其中包含靶标功能、序列、三维结构、配体结合特性、酶命名法和药物结构、治疗类别、临床发展状况等信息[25]。在网络药理学研究中，TTD 数据库中包含的已知药物结构和靶标信息可作为未知药物靶标预测的阳性对照数据集，通过化合物结构和功能相似性的比对，获得未知药物的候选靶标谱。

1. 数据结构

为了解针对不同靶标的不同分子支架而开发的 QSAR 模型对于促进药物开发和优化工作非常有用。目前，针对 121 种靶标和 228 种活性化合物，TTD 构建了 841 种基于配体的 QSAR 模型，并可在相关页面访问和了解模型的具体架构。

2. 功能介绍

随着生物信息学的快速发展，数据库技术在生物信息学中发挥了重要作用。药物研究的困难在于靶点的发现和确定，TTD 收集三种类型的验证数据：通过实验确定药物对其主要靶点的效力，观察药物对与其主要靶点相关的疾病模型（细胞系、体外、体内模型）的效力或效果，以及目标敲除、RNA 干扰、转基因等治疗的体内模型的观察效果。

3. 特性

TTD 数据库架构和接口设计，便于用户访问更新数据和以往版本数据。使用 Drupal 作为增强数据存储和提取的数据库平台。在新的 TTD 界面中，可以通过患者数据手册栏访问新添加的耐药性突变和目标表达数据，并且可以通过目标药物手册栏访问目标组合信息。Drugs Group 手册栏还包括多靶向制剂和自然衍生药物搜索选项。高级搜索手动栏包括自定义搜索、目标相似性搜索、药物相似性搜索和路径搜索选项。此外，添加了 JSME 分子编辑器，以方便用户绘制分子并随后搜索结构与输入分子相似的 TTD 药物条目。目前，该数据库还在不断地更新完善中，收录越来越多的靶点信息不仅扩大了靶点的信息量，对于药物新靶点发现、药物筛选、疾病治疗以及药理学机制的研究均具有重大意义。

3.3.2　PDB：蛋白质晶体结构数据库

PDB（Protein Data Bank）数据库是一个生物大分子结构数据库，最初由美国 Brookhaven 国家实验室的 Walter Hamilton 博士于 1971 年建立，并于 1973 年正式向全世界有关实验室提供数据[26]。1998 年 10 月，PDB 被移交给了结构生物信息学研究合作组织（RCSB），并于 1999 年 6 月移交完毕。直到现在，PDB 数据库的维护都由 RCSB 负责。RCSB 的主服务器和世界各地的镜像服务器提供数据库的检索和下载服务。PDB 收集了通过实验（X 射线晶体衍射、核磁共振、电子显微等方法）测定的 153 085 个生物大分子的三维结构数据，主要是蛋白质，还包括核酸、多糖、蛋白质与核酸复合物和各类由 X 射线晶体衍射、核磁共振分析方法测定的合成物。无论是通过哪种途径获得的结构数据，在 PDB 中均以相同的格式存储在一个空间结构数据库中，称为一个 Entry。每一个 Entry 都有其唯一的 PDB-ID，由 4 个字符（大写字母 A～Z 和数字 0～9 中的 4 个）组合而成，如：6A21。用户可以通过输入 PDB-ID 在 PDB 中查询到相关信息，包括分子名、该分子的收录日期、样品来源、作者姓名、ID 号、序列、一级结构、二级结构（α-螺旋，β-折叠及 β-转角）、异质（对非标准氨基酸残基的说明）、连接部分（二硫键及其他一些化学连接情况）、原子的空间坐标及末端组成、测定结构所用的实验方法、衍射数据的分辨率、相关文献等信息，如图 3-22 所示[27]。PDB 中所有的数据都可以通过网络免费访问，还可以从发行的光盘获得，为网络靶标的结构解析和功能挖掘提供数据支持。

1. 数据结构

PDB 中生物大分子的结构是通过特定的格式，以原子空间坐标值和对于其连接形式、连接顺序等的描述来表示的。通过特定的软件，如 PyMol、RasMol、Chimera、VMD、Swiss-PdbViewer 等，在计算机上按 PDB 文件实现生物大分子的三维立体结构可视化，对结构进行详细的查看、编辑，从而应用于进一步的研究。

用户通过 PDB 基于万维网的 AutoDep 设施，以 mmCIF 或 PDB 的格式向 PDB 提交数据（如图 3-23 所

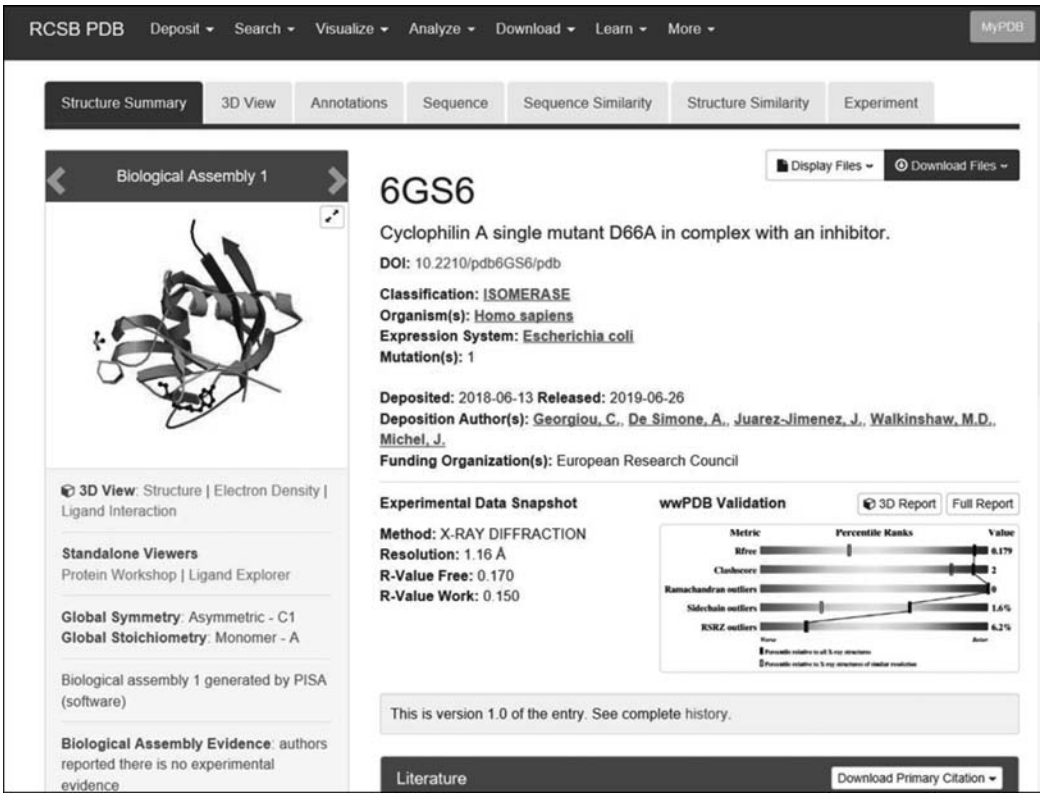


图 3-22 PDB 中 6GS6 查询结果截图

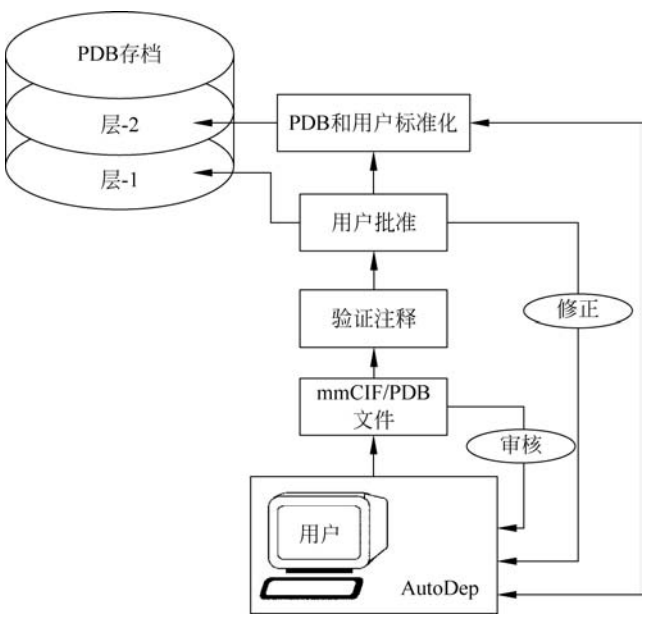


图 3-23 基于万维网 AutoDep 的分层提交方法

示)。然后 AutoDep 调用一套验证程序,在数据发送到 PDB 后几分钟内通过万维网将输出的诊断文件返回给用户。其中通过验证的条目将被标识为 LAYER-1(层-1)发布。然后,PDB 工作人员对需要验证的条目和输出文件进行评估后,完成注释并返回给用户以征求意见和批准。在用户批准和进行更正后,最终的条目将被指定为 LAYER-2(层-2)发布。

PDB 中的三维结构记录,可分为两种:显性序列信息和隐性序列信息。两者都可用于重构生物高聚体的化学图像。显性序列在 PDB 文件中以关键词 SEQRES 开头逐行存储。不同于其他序列数据库,PDB 记录使用三字母氨基酸编码(如图 3-24 所示)。PDB 记录中的隐性序列,即为立体化学信息,蕴涵在 PDB 文件中的 ATOM 记录及相应的(X,Y,Z)三维坐标立体结构中。ATOM 记录列出了每一原子的原子名,所属残基名,残基顺序号,原子的 X、Y、Z 坐标,占有率及温度因子等信息(如图 3-25 所示),约占每一数据文件中全部记录总数的 90%以上[28]。

```
SEQRES   1 A  273  MET ARG GLN ILE ALA ILE TYR GLY LYS GLY GLY ILE GLY
SEQRES   2 A  273  LYS SER THR THR THR GLN ASN LEU THR ALA ALA LEU SER
SEQRES   3 A  273  THR MET GLY ASN ASN ILE LEU LEU VAL GLY CYS ASP PRO
SEQRES   4 A  273  LYS ALA ASP SER THR ARG MET LEU LEU GLY GLY LEU ASN
SEQRES   5 A  273  GLN LYS THR VAL LEU ASP THR LEU ARG SER GLU GLY ASP
SEQRES   6 A  273  GLU GLY ILE ASP LEU ASP THR VAL LEU GLN PRO GLY PHE
SEQRES   7 A  273  GLY GLY ILE LYS CYS VAL GLU SER GLY GLY PRO GLU PRO
SEQRES   8 A  273  GLY VAL GLY CYS ALA GLY ARG GLY ILE ILE THR SER ILE
SEQRES   9 A  273  GLY LEU LEU GLU ASN LEU GLY ALA TYR THR ASP ASP LEU
SEQRES  10 A  273  ASP TYR VAL PHE TYR ASP VAL LEU GLY ASP VAL VAL CYS
SEQRES  11 A  273  GLY GLY PHE ALA MET PRO ILE ARG GLU GLY LYS ALA LYS
SEQRES  12 A  273  GLU ILE TYR ILE VAL ALA SER GLY GLU LEU MET ALA ILE
SEQRES  13 A  273  TYR ALA ALA ASN ASN ILE CYS LYS GLY LEU ALA LYS PHE
SEQRES  14 A  273  ALA LYS GLY GLY ALA ARG LEU GLY GLY ILE ILE CYS ASN
SEQRES  15 A  273  SER ARG LYS VAL ASP GLY GLU ARG GLU LEU LEU GLU ALA
SEQRES  16 A  273  PHE ALA LYS LYS LEU GLY SER HIS LEU ILE HIS PHE VAL
SEQRES  17 A  273  PRO ARG ASP ASN ILE VAL GLN ARG ALA GLU ILE ASN ARG
SEQRES  18 A  273  LYS THR VAL ILE ASP PHE ASP ARG GLU SER ASP GLN ALA
SEQRES  19 A  273  LYS GLU TYR LEU THR LEU ALA ASP ASN VAL GLN ASN ASN
SEQRES  20 A  273  ASN LYS LEU VAL VAL PRO THR PRO LEU PRO MET GLU GLU
SEQRES  21 A  273  LEU GLU ALA MET MET VAL GLU PHE GLY ILE VAL GLU LEU
SEQRES   1 B  273  MET ARG GLN ILE ALA ILE TYR GLY LYS GLY GLY ILE GLY
SEQRES   2 B  273  LYS SER THR THR THR GLN ASN LEU THR ALA ALA LEU SER
SEQRES   3 B  273  THR MET GLY ASN ASN ILE LEU LEU VAL GLY CYS ASP PRO
SEQRES   4 B  273  LYS ALA ASP SER THR ARG MET LEU LEU GLY GLY LEU ASN
```

图 3-24　6QA8 的 PDB 文件显性序列部分截图

2. 功能介绍

1) 蛋白质二级结构的预测

有些研究者利用 PDB 中已搜集的蛋白质二级结构信息,试图归纳出更可靠的二级结构预测方法,已有很多结果发表。随着技术的不断进步和越来越多成果的出现,也许在不久的将来会有所突破。

2) 蛋白质进化的研究

以前主要是从一级结构出发,分析序列的异同。随着 PDB 中蛋白质立体结构信息的大量增加,目前,某些研究人员正试图从立体结构的结构相关性出发来研究蛋白质进化问题。

3) 模拟蛋白质的卷曲过程

研究人员提出了各种可能的卷曲路径,并在理论上模拟了卷曲过程,但这些假设的正确性,必须与已经明确的蛋白质空间结构比较才能确定。PDB 就提供了这样的一个比较标准,使研究人员可根据比

```
ATOM      1  N   MET A   1     -27.296  10.466  -9.085  1.00 60.81           N
ATOM      2  CA  MET A   1     -27.129  11.862  -8.690  1.00 58.02           C
ATOM      3  C   MET A   1     -28.472  12.594  -8.599  1.00 51.30           C
ATOM      4  O   MET A   1     -29.247  12.587  -9.549  1.00 54.67           O
ATOM      5  CB  MET A   1     -26.209  12.589  -9.676  1.00 45.14           C
ATOM      6  CG  MET A   1     -25.956  14.046  -9.312  1.00 50.94           C
ATOM      7  SD  MET A   1     -25.155  15.037 -10.600  1.00 60.91           S
ATOM      8  CE  MET A   1     -26.378  14.949 -11.908  1.00 57.37           C
ATOM      9  N   ARG A   2     -28.747  13.225  -7.459  1.00 48.49           N
ATOM     10  CA  ARG A   2     -29.942  14.053  -7.342  1.00 50.60           C
ATOM     11  C   ARG A   2     -29.745  15.367  -8.092  1.00 51.85           C
ATOM     12  O   ARG A   2     -28.657  15.957  -8.073  1.00 48.99           O
ATOM     13  CB  ARG A   2     -30.268  14.335  -5.873  1.00 48.21           C
ATOM     14  CG  ARG A   2     -30.578  13.104  -5.033  1.00 47.54           C
ATOM     15  CD  ARG A   2     -32.058  12.706  -5.111  1.00 49.73           C
ATOM     16  NE  ARG A   2     -32.934  13.548  -4.293  1.00 47.36           N
ATOM     17  CZ  ARG A   2     -33.068  13.440  -2.972  1.00 49.01           C
ATOM     18  NH1 ARG A   2     -32.372  12.531  -2.303  1.00 48.83           N
ATOM     19  NH2 ARG A   2     -33.902  14.242  -2.313  1.00 50.05           N
ATOM     20  N   GLN A   3     -30.808  15.830  -8.749  1.00 45.05           N
ATOM     21  CA  GLN A   3     -30.766  17.026  -9.590  1.00 44.03           C
ATOM     22  C   GLN A   3     -31.820  18.016  -9.095  1.00 45.48           C
ATOM     23  O   GLN A   3     -32.992  17.957  -9.479  1.00 40.95           O
ATOM     24  CB  GLN A   3     -30.975  16.651 -11.040  1.00 46.05           C
ATOM     25  CG  GLN A   3     -30.129  15.472 -11.466  1.00 45.26           C
ATOM     26  CD  GLN A   3     -30.345  15.109 -12.911  1.00 51.08           C
```

图 3-25　6QA8 的 PDB 文件隐性序列部分截图

较结果有目的地修改假设，使之更符合实际情况。

3. 特性与不足

与其他关于分子立体结构的数据库不同，PDB 搜集的数据绝大部分未经公开发表，是直接由各实验室向 PDB 提供的；而其他数据库中的数据则主要来自于公开的出版物[29]。

尽管 PDB 的处理程序已经有了很大的改进，但仍然存在许多不被系统发现的错误，需要对所有条目进行人工检查。且某些类型的问题仍然需要人工干预和处理，如：涉及处理异质（结构复杂的小分子）和解决晶体堆积的问题，提交的氨基酸序列和数据库中查到的氨基酸序列之间存在冲突等问题。有时需要参考其他的出版物和资料来明确事实信息，如晶体数据、生物细节等。需要进一步扩展和完善 AutoDep 的录入和验证程序组件，以适应储户和用户之间有些冲突的需求，同时确保信息保持最高标准的准确性。

3.3.3　GeneCards：基因信息数据库

GeneCards 数据库基本覆盖了各种专业数据库对人类基因的分析数据[30]，是一个较为全面、好用的人类基因组注释数据库。GeneCards 数据库于 1997 年由以色列的 Weizmann 科学研究所的皇冠人类基因组中心建立，该数据库建立之初目标是将分散在各种数据库中的片段信息合理地、系统地整合到一起。通过二十多年的研发和维护，GeneCards 数据库克服了不同数据库自身格式的局限，自动挖掘并集成了 190 余个数据库的成千上万个人类基因表达、功能、位置、通路、变异、同源基因、疾病和相关参考文献等各种信息整合成基因网络卡片以供研究人员参考应用。截至 2019 年 11 月，GeneCards 数据库已收录并整理了 268 549 个人类基因的数据。该数据库中的数据不仅及时更新，还可以免费浏览。

GeneCards 数据库作为一种人类基因纲要，包括基因组、转录组、蛋白质组、遗传、临床、功能等以基

因为中心的信息[31]。为了使数据库页面更加紧凑，GeneCards 数据库中部分条目的详细内容、图、表及参考文献等通过单击超链接形式即可查看到相关基因的所有可用信息。

1. 数据结构

GeneCards 数据库中的每个基因条目都以电子化的网页卡片形式分为 17 个主要章节进行描述。GeneCards 数据库中每个基因卡片通过编制各种数据库中的基因信息，定期自动整合并注释原数据库基因信息到该数据库中的相应章节。GeneCards 数据库自改版到 4.0 版本以后在保留传统内容与功能的基础上，通过整合数据与信息达到更好地用户体验。

2. 在分子生物领域的应用/功能

GeneCards 数据库中对每个基因的描述非常详细。基因功能方面，①分子功能：GeneCards 数据库中分子功能来源于其他数据库，基因本体分子功能表中展示了 GO IDs、GO 术语、证据和 PubMed IDs，此外超链接允许用户查看其他共享这个基因本体的基因。②表型和动物模型：此部分列出了人和小鼠基因相关表型，超链接不仅可以允许用户查看到其他共享这个基因表型的基因，还可以导航至动物模型，如敲除该基因的鼠模型。③功能相关产品：此部分内容提供了与该基因相关动物模型、克隆和细胞系等产品链接。通路与相互作用方面，①超级代谢途径：超级代谢途径表格中展示了该基因可能涉及的通路与动态链接，其中 G 为 KEGG 通路信息，R 为 REACTOME 通路信息。②通路来源：此部分内容按照通路信息来源数据库进行分类，每个数据库单元中仅显示 5 条相关通路，通过单击上方链接可查看全部通路信息。③相互作用蛋白质：蛋白质相互作用网络以图片形式展示，单击链接可查看更复杂、更多相互作用信息的网络图片。此外，以表格形式列出了相互作用蛋白质信息，包括每个相互作用蛋白质的基因名称与基因卡片链接、相互作用蛋白质的 ID 与外部数据库的链接，以及蛋白质相互作用网络中的相互作用信息链接。④相互作用信号网络开放资源（the Signaling Network Open Resource，SIGNOR）：呈现了相互作用信号网络开放资源链接，以及相互作用基因列表与基因卡片链接。⑤基因本体生物过程：此部分展示了基因本体生物过程，包括该基因 GO ID、GO 术语、证据和 PubMed ID[32]。

3. 特性与不足

GeneCards 数据库经过二十多年的发展与数十次的改版，目前该数据库收录了 190 余个数据库的 268 549 个人类基因表达、功能、位置、通路、变异、同源基因、疾病和相关参考文献等各种信息，并整合成基因网络卡片供研究人员参考应用，是集成了多种专业数据库功能的人类基因综合数据库。

GeneCards 数据库中基因信息虽然丰富多样，但为了使数据库网页更加紧凑，数据中部分条目的详细内容、图、表及参考文献等通过单击超链接形式可查看到相关基因的所有可用信息，这就意味着用户需要频繁单击超链接或导航至其他数据库中获取基因的全面信息，间接地浪费了用户的研究时间。

3.3.4　KEGG：京都基因与基因组百科全书

如何借助计算机全面地展示细胞和生物所包含的生物学信息是后基因组时代的重大挑战之一。科学家们期望能够根据基因组中的信息，计算或者预测复杂的细胞通路或者生物学反应。为此，日本京都大

学生物信息学中心的 Kanehisa 实验室于 1995 年建立了京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes),即 KEGG 数据库。该数据库最新更新时间为 2022 年 1 月 1 日,最新发布的版本为 101.0。KEGG 是一个集成的数据库资源,分为系统信息、基因组信息、化学信息和健康信息等。KEGG 将基因、基因组信息以及更高层次的功能信息结合起来,通过对细胞内已知生物学过程的计算机化和将现有的基因功能信息解释标准化,对基因的功能进行系统化分析[17],且具有描述代谢途径预测基因功能获取基因组信息同源性识别以及解析蛋白质和其他大分子相互作用等诸多功能。研究人员不仅可以免费获取该数据库的数据,还可以使用 Java 图形工具访问基因组图谱,比较基因组图谱和转录组表达图谱。

1. 数据结构

迄今,KEGG 数据库共有 18 个子数据库,其中 4 个主要数据库为 PATHWAY、GENES、LIGAND、BRITE,其他子数据库是在这 4 个数据库基础上衍生而来的。PATHWAY 数据库提供发生在细胞内各种反应的人工绘制途径图,以网络形式呈现。GENES 数据库储存 KEGG 中注册的已测序的基因组信息。LIGAND 数据库可用于查询化合物、多糖及酶促反应等信息。BRITE 是将生物信息按等级层次分类归纳的数据库,其中所包含的 KEGG ORTHOLOGY(KO)是用于基因同源性识别的系统[18]。

2. 功能介绍

KEGG 作为一个参考知识库,被广泛地用于基因组测序和其他高通量实验技术得到的大规模数据集的整合和解释[19]。其在生物信息学中的应用包括:①代谢网络的分析。KEGG 通路图、BRITE 分层条目和 KEGG modules 构成了 KEGG 参考信息。用 KEGG mapper 来标记通路,就可以对代谢通路中需要的化合物或酶着色显示,有利于代谢途径的分析。另外,还可以对基因芯片数据进行分析,例如,在 KEGG Expression 数据库中分析基因芯片数据时,KegArray 可以使用不同颜色表示通路中各基因表达的变化,红色表示上调,绿色表示下调。②疾病及药物代谢网络分析应用。KEGG Mapping 整合疾病和药物信息广泛用于相关研究。收集在 KEGG DISEASE 的所有已知疾病基因以及收集在 KEGG DRUG 的所有药物靶点都合并在 KEGG PATHWAY 和 BRITE 数据库中,可以在代谢图中使用 KEGG Mapping 用不同颜色标出对应基因。在疾病的代谢路径图里的疾病/药物图中,粉色框里是与疾病有关基因,亮蓝色框里是药物靶点。③基因组比较以及合并。在 KEGG GENOME 页面不仅可以用 Mapping 比较不同物种的代谢能力,还可以来检查人-病原体以及人-微生物代谢关系互补性,检查物种之间的共同特征。④重构代谢网络以及目标物种酶数据库的构建。从 LIGAND 数据库中能够获取重建目标物种的代谢网络中的所有基因-酶以及酶-反应列表,其中,酶在连接基因和相应代谢反应中起到关键作用,由于酶的 EC 号是唯一的,可以据此建立一个包含参与细胞新陈代谢的所有代谢组分及其代谢反应的列表。再通过其他数据库的信息辅助参考优化,就可以构建出该目标物种全部酶及反应数据库。得到高质量数据库后,即可用相关软件对代谢网络进行重构[20]。因此,KEGG 数据库可被广泛运用于代谢网络的构建。

3. 特性与不足

KEGG 数据库作为一个联系了基因、酶和反应构建代谢网络的大型综合数据库,它以善于分析解读的图形界面为独特优势,对基因、酶以及其代谢网络的研究提供了很好的平台。在生物合成方面,能够通过控制代谢流量来提高目标产物产量。不仅如此,KEGG 着重发展的疾病代谢网络有助于研究疾

病致病机理以及药物作用靶点。

KECG 是一个较全面的数据库，包含了代谢通路、基因信息、化合物反应等数据，但是也有一些疏忽之处，比如，着色输入框区分大小写；KegArray 启动时数据不对；有些酶促反应在 LIGAND 库中虽有记录但在指定物种中并不发生，因而导致重构的网络包含虚假边；在重构多个代谢网络的情况下，为读取数据不得不频繁访问 KEGG 的远程服务器，非常耗时等问题。

3.4　网络药理学常用蛋白相互作用数据库

3.4.1　BioGRID：生物学通用相互作用数据库

BioGRID(Biological General Repository for Interaction Database)数据库是一个免费开放的交互存储库，致力于蛋白质信息的管理和储存、所有生物物种及人类的遗传物质和化学相互作用。BioGRID 创建于 2003 年，最初为一般的交互数据集存储库，后更名为 BioGRID[34]。当前 BioGRID 的版本为 3.5.173。该版本从 69 644 篇文献中整理出了 1 690 901 个蛋白质-基因相互作用信息，与 28 093 个化学关联以及 726 378 个蛋白转录后修饰信息，并覆盖动物类(包括人、斑马鱼、小鼠、果蝇、冈比亚按蚊、欧洲蜜蜂、牛、犬、豚鼠、猴、马、鸡、兔、羊、猩猩、猪、非洲爪蟾、海胆等)，植物类(包括拟南芥、金银花、大豆、稻、番茄、马铃薯、葡萄、玉米、蓖麻等)，细菌病毒类(包括枯草杆菌、白色念珠菌、大肠杆菌、肝炎病毒、疱疹病毒、HIV 病毒、人乳头瘤病毒、结核分枝杆菌、粗糙脉胞霉、烟草花叶病毒、玉米黑粉菌、牛痘病毒)，其他包括秀丽隐杆线虫、阿米巴虫、虱子、疟原虫、酵母、地衣类等。所有信息都可以通过网站提供的搜索引擎免费查看和下载，数据库还提供了多个在线分析和可视化工具。

BioGRID 网站的主页如图 3-26 所示，使用起来也比较简单，只需要输入一个基因 ID、关键词或基因名，选择物种，单击搜索即可获得基因互作的结果。检索结果主要由三部分内容组成。①基本信息的描述：包括检索的蛋白质名词、别名，转录后修饰，GO 注释信息以及和其他数据库的链接；②信息统计：统计每种互作用类型和比例；③详细结果显示：提供蛋白相互作用信息、该蛋白质的相互作用网络等。

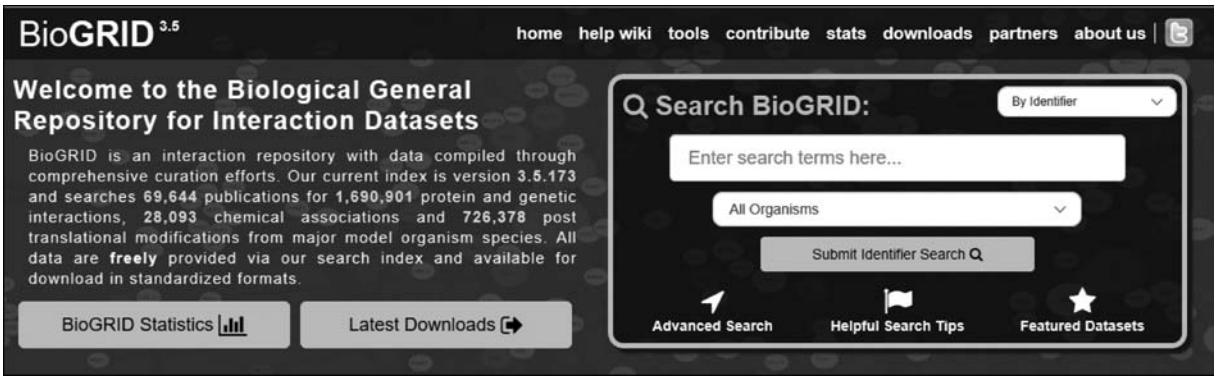


图 3-26　BioGRID 网站的主页

1. 数据结构

BioGRID 还在生物医学科学特别相关的领域进行相关项目开发，例如泛素-蛋白酶体系统和各种人

类疾病相关的相互作用网络。BioGRID 策略通过交互管理系统(Interactive Multimedia Service,IMS)进行协调,该系统通过结构化证据代码,表型本体和基因注释促进编译交互记录。BioGRID 结构已得到改进,以支持更广泛的相互作用和翻译后修饰类型,允许更复杂的多基因/蛋白质相互作用的表示。

2. 功能介绍

随着后基因组时代的发展,蛋白质研究越来越广泛和深入。BioGRID 持续扩大从生物医学文献中筛选蛋白质和遗传相互作用,以及相关属性,如蛋白质变异、表型和化学或药物相互作用等。这些网络数据集与其他数据类型(包括表达数据、定量表型数据和高分辨率序列数据)的集成将推动药物发现研究工作。

3. 特性

当前侧重于生物学的特定领域,目前正在努力扩大对多种后生动物的管理,以便深入了解与人类健康相关的保守网络和路径。BioGRID 3.5 Web 界面包含新的搜索和显示功能,可以跨多种数据类型和来源进行快速查询。BioGRID 为几个模型生物数据库提供交互数据,如 Entrez-Gene,SGD,TAIR 等资源,FlyBase 和其他交互元数据库。整个 BioGRID 3.2 数据集可以以多种文件格式下载,包括 IMEx 兼容的 PSI MI XML。对于开发人员,BioGRID 交互也可通过基于 REST 的 Web 服务和 Cytoscape 插件获得。所有 BioGRID 文档均可在 BioGRID Wiki 中在线获取。

3.4.2 DIP:蛋白质相互作用数据库

DIP(Database of Interacting Proteins)数据库是 1999 年 8 月由加州大学洛杉矶分校分子生物学研究所的结构生物学和分子医学实验室创立的,旨在将蛋白质相互作用(Protein-Protein Interaction,PPI)的各种实验证据整合到一个易于访问的在线数据库中,建立一个简单、易用的 PPI 公共数据库。此外,DIP 数据库还是 IMEx 联盟(International Molecular Exchange Consortium)的成员数据库之一。

1. 数据结构

DIP 数据库收录了经实验证实的 PPI 信息以及来自 PDB(Protein Data Bank)数据库的蛋白质复合物,属于经过专家手工挖掘或通过计算方法获得最可靠的 PPI 数据[35-38]。截至最近更新时间 2017 年 2 月 13 日,DIP 收集归类了涵盖 834 种物种来源的 28 850 种蛋白参与的 81 923 个 PPI,覆盖了 8234 个数据源共 82 143 项实验。

DIP 数据库提供多种查询方式,用户可直接基于蛋白质、生物物种、蛋白质超家族、关键词、实验技术或引用文献查询 PPI,也可基于序列相似性的 BLAST 搜索、pattern 搜索和 motif 搜索查询 PPI。查询的结果列出节点(Node)与连接(Link)两项,节点用于描述所查询的蛋白质的特性,包括蛋白质的功能域(Domain)、指纹(Fingerprint)等,有的还批注酶的代码或出现在细胞中的位置;连接指两个节点之间的相互作用关系。DIP 对每一个 PPI 都会说明证据(实验的方法)、提供文献,也记录除巨量分析外的支持此 PPI 的实验数量。查询结果的示例如图 3-27 所示。DIP 数据库提供的标准数据集包括 HiTHR 高通量(基因组规模)数据集、FULL(完整的 DIP 数据集)、SPECIES(特定物种集)、FASTA(DIP 序列)等和 DIP-IMEx 数据集[39,40]。

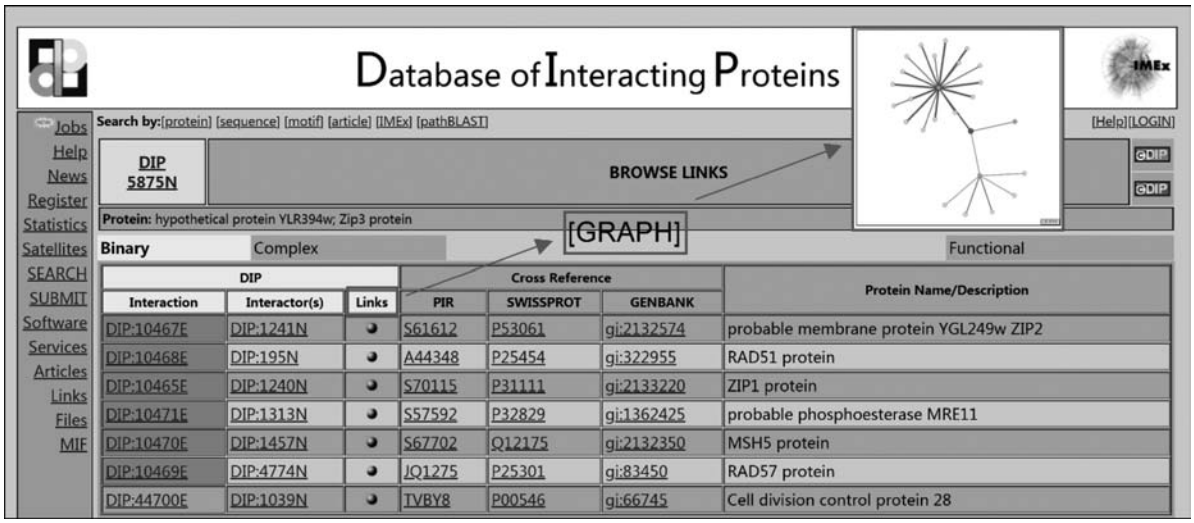


图 3-27　DIP 数据库查询结果示例

2. 功能介绍

MiSink 是 DIP 数据库用于 Cytoscape(一个用于生物交互数据可视化和集成的开源平台)的一个插件,可将其转换为 DIP 的交互式图形界面。JDIP 是 DIP 数据库提供的一个基于 Java 语言的可视化应用工具,可将 PPI 数据以网络形式更加直观地展现出来,并允许用户将其实验信息如 mRNA 表达数据、功能结构域的功能、蛋白质翻译后修饰等整合到蛋白质之间相互作用的网络中。此外,DIP 数据库还发展了 3 个子数据库[35,40]:蛋白质配体与受体数据库(The Database of Ligand-Receptor Partners,DLRP),实时 PPI 数据库(The LiveDIP Database,LiveDIP),用基因融合法、系统发生谱法等预测的 PPI 数据库(Inferring Functional Linkages Between Proteins,Prolink)。

3. 特性

DIP 数据均为经过专家手工挖掘或通过计算方法获得最可靠的 PPI 数据,它被选为评估高通量筛选和计算机预测得到的 PPIs 的黄金标准集,并且提供 PPIs 真实性评估服务,包括基于平行同源关系的 PVM(Paralogous Verification Method)、基于表达谱分析的 EPR 法(Expression Profile Reliability)以及基于结构域相互作用的 DPV 法(Domain Pair Verification)。但相应地,必要的人工干预和处理使该数据库更新相对缓慢。

3.4.3　IntAct:分子相互作用数据库

IntAct(Molecular Interaction Database)数据库是由欧洲生物信息学研究所于 2003 年创立[41,42]的,主要目标是帮助研究人员充分利用公共 PPI 数据,减少冗余并提供一个统一的查询工具,最大化提升数据存储和检索的效率。IntAct 数据库是 IMEx 联盟的成员数据库之一,目前该数据也整合了 IMEx 的所有数据。IntAct 数据库为分子间相互作用提供了一个免费的、有开源数据库的分析工具,所有的数据来源于已经发表的文献报道结果,并由生物学专家人工注释,保证高精确性,包括实验方法、实验条件和相互作用的功能结构域等[42]。IntAct 数据库的最新版本为 4.2.17,涵盖物种有人、酵母、果蝇、大肠杆菌、拟南芥(鼠耳藓)和秀丽隐杆线虫,包含了 110 643 个蛋白或分子的 585 731 个相互

作用和 889 774 个二元相互作用的证据，覆盖了 20 585 个出版物、67 624 项实验，共有 3829 个受控词(Controlled Vocabularies)对用于生成数据的实验细节进行一致的描述。

IntAct 数据库分基本查询和高级查询，基本查询可根据基因名称、蛋白质名称、PubMed ID 和生物学作用等进行简单搜索；高级查询可根据实验方法和 IntAct 自定义的受控词进行查询，结果展示工具可以显示图形化的 PPI 网络。

1. 数据结构

IntAct 数据库支持包括 PSI-MI XML，PSI-MITAB，RDF/XML，RDF/XML-ABBREV，N3，N-Triples 和 Turtle 在内的多种格式。IntAct 研究小组建议生物学家在文献发表之前向该数据库直接提交 PPI 信息(格式不限，推荐 IMEx 格式)，这一过程如同向 GenBank 数据库直接提交核苷酸序列一样，可以方便数据的增加和管理。IntAct 数据可通过 PSICQUIC 服务以及许多其他数据类型获得，包括预测性交互，基因组和用于推断分子相互作用的基于文本挖掘方法的结果。

2. 功能介绍

IntAct 数据库提供 PPI 网络的可视化在线分析，同时支持 Cytoscape，Proviz 等第三方网络构建软件。除了存储并查询相互作用蛋白质信息，IntAct 数据库还提供基于“Pay-As-You-Go”算法预测下拉实验(Pull-Down)的最佳诱饵蛋白信息。

3.4.4 STRING：基因/蛋白相互作用关系数据库

STRING(Search tool for the retrival of interacting genes/proteins)数据库由欧洲分子生物学实验室于 2009 年创建[43~46]，旨在收集和整合已知和预测的大量生物蛋白质-蛋白质关联数据信息。它是一个免费、开源的 PPI 检索与预测信息数据库，整合了来源于高通量实验、文本挖掘、生物信息预测和相互作用数据库(如 BioGRID 和 IntAct 等)的 PPI，同时利用打分系统对不同方法得到的相互作用分配不同权重，提供每对 PPI 的可靠性评分[35,36]。目前最新版本为 2019 年 1 月 19 日发布的 11.0 版，涵盖 5090 种有机体的约 24 600 000 个蛋白和超过 2 000 000 000 个相互作用。

1. 数据结构

STRING 数据库中的关联包括具有特定性和生物学意义的直接(物理)相互作用以及间接(功能)相互作用，同时利用打分系统对不同方法得到的相互作用分配不同权重，提供每对 PPI 的可靠性评分。除收集和重新评估 PPI 的现有数据(来源有：KEGG，EcoCyc，BIioCyc，GO，Reactome，Biocarta，NCI-Nature Pathway Interaction Database，MINT，HPRD，BIND，DIP，PID，BioGRID)，以及用数据集导入已知途径和蛋白质复合物外，相互作用预测有以下来源：①系统共表达分析；②共有选择性检测跨基因组的信号；③科学文献的自动文本挖掘(来源有：SGD，OMIM，FlyBase，PubMed)；④基于基因直系学的生物之间相互作用知识的计算转移(Neighborhood，Co-occurrence，Co-expression，Gene Fusion)。

用户可根据蛋白质名称(可同时输入多个)、序列名称(可同时输入多个)、生物体或蛋白质家族进行查询，结果以由节点(Node)和边(Edge)组成的可单击的互动网络图进行展示，节点表示蛋白，节点之间的连线表示两个蛋白之间的相互作用，也可根据需要选取特定来源的数据或扩展的网络图进行重新绘图。结果的导航选项包括 Viewers，Legend，Settings，Analysis，Exports，Clusters 和用于调整互动网络

图中显示节点数量的 More/Less，选中一个节点处的蛋白，可在弹出窗口中显示其结晶蛋白（来自 PDB）的图像以及蛋白质模型（来自 SwissModel）的图像等，并允许进行以下操作：①查找 STRING 中与窗口蛋白质相互作用的所有蛋白质；②向网络添加与窗口蛋白质相互作用的蛋白质；③显示蛋白质序列；④STRING 中的同系物；⑤重定向到 GeneCards 数据库中的相应条目（仅适用于人类蛋白质）；⑥重定向到 SMART 数据库中的相应条目。在 Viewers 页面中，用户可获得 Network，Neighborhood，Co-occurrence，Co-expression，Fusion，Experiments，Databases 和 Textmining 等相关信息。在 Legend 页面中，显示了每个蛋白的颜色和对应的与查询 PPI 的 score 值。在 Settings 页面中，用户可对结果中的 PPI 类型和呈现方式进行设置。在 Analysis 页面中，对于 PPI 网络中的基因，提供了 GO 和 KEGG 富集分析的结果。在 Clusters 页面中，用户可对基因进行聚类分析，支持 K-Means 和 MCL 聚类，聚类的结果为 TSV 格式。查询结果示例如图 3-28 所示。

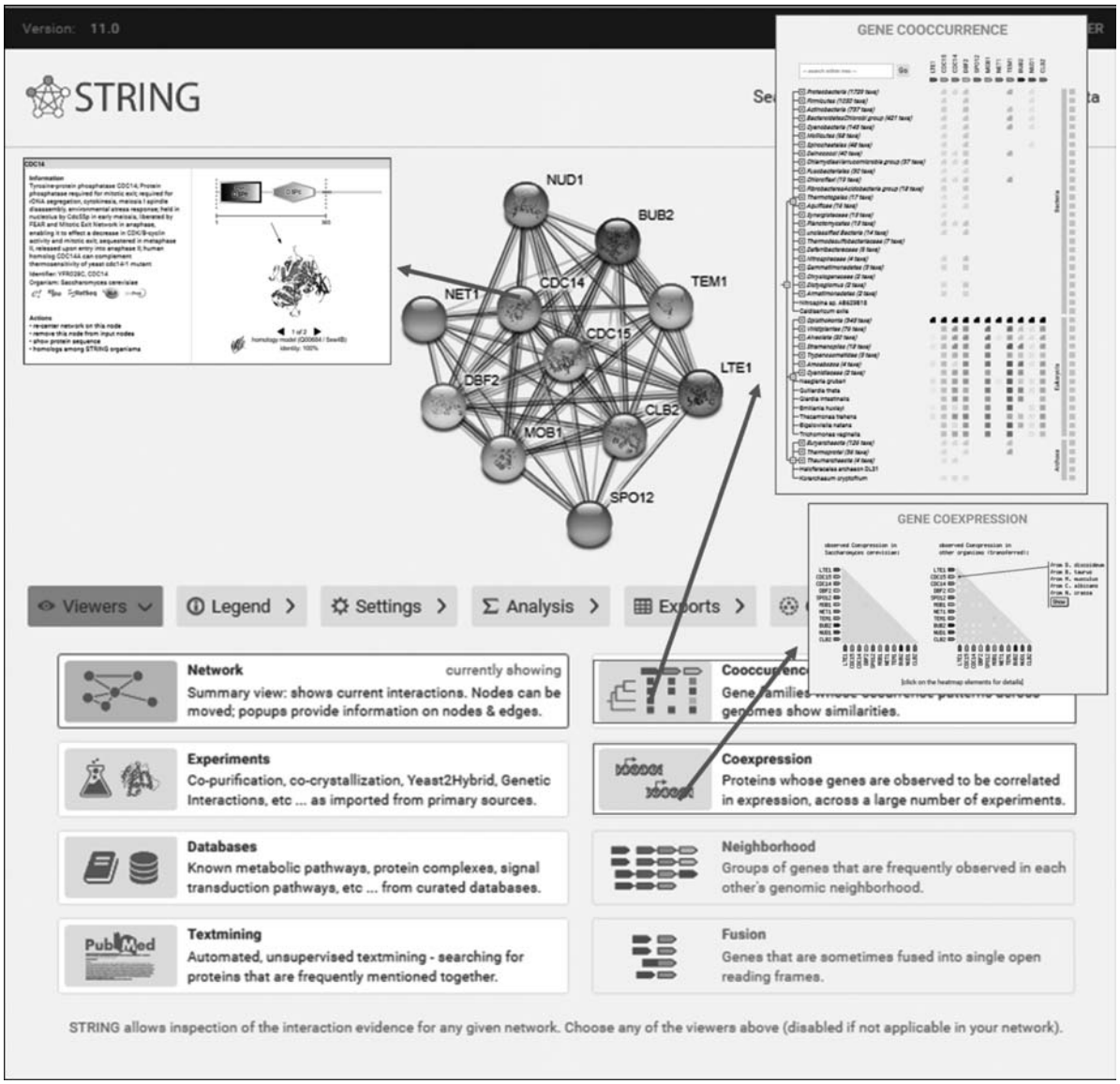


图 3-28　STRING 数据库查询结果示例

2. 功能介绍

STRING 数据库的主要目的是构建 PPI 网络。它可以用于过滤和评估功能性基因组学的数据，并为注释蛋白质的结构、功能和进化性提供一个比较直观的平台。用于探索预测 PPI 网络，为实验研究提供新方向，并且能够为相互作用的映射提供物跨物种预测。所有的 PPI 数据都被加权、整合，并且都有一个计算得到的可靠值。

3. 特性

STRING 数据库完全是预先计算好的，因此高层次的网络或单个 PPI 的界面的所有信息都可以被迅速获取。支持单独选择各种证据类型，可在运行的时候进行定制搜索，同时也会有专门的查看器来对所有的关联证据进行查看。STRING 数据库是一项探索性的资源，比基本的 PPI 数据库包含了更多的关联数据。推荐用于快速、初步地获取要查询的蛋白的 PPI 信息，尤其是未有良好表征的蛋白。

参 考 文 献

[1] XU H Y, ZHANG Y Q, LIU Z M, et al. ETCM: an encyclopaedia of traditional Chinese medicine[J]. Nucleic Acids Research, 2019, 47(D1): 976-982.

[2] HUANG L, XIE D L, YU Y R, et al. TCMID 2.0: a comprehensive resource for TCM[J]. Nucleic Acids Research, 2018, 46(D1): 1117-1120.

[3] XUE R C, FANG Z, ZHANG M X, et al. TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis[J]. Nucleic Acids Research, 2013, 41(D1): 1089-1095.

[4] WU Y, ZHANG F L, YANG K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping[J]. Nucleic Acids Research, 2019, 47(D1): 1110-1117.

[5] FANG Y C, HUANG H C, CHEN H H, et al. TCMGeneDIT: a data-base for associated traditional Chinese medicine, gene and disease information using text mining[J]. BMC Complementary and Alternative Medicine, 2008, 8(1): 58.

[6] RU J L, LI P, WANG J N, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines[J]. Journal of Cheminformatics, 2014, 6(1): 13.

[7] LIU Z Y, GUO F F, WANG Y, et al. BATMAN-TCM: a bioinformatics analy-sis tool for molecular mechanism of traditional chinese medicine[J]. Scientific Reports, 2016, 6(1): 82-83.

[8] YOAV B, YOSEF H. Controlling the false discovery rate: a practical and powerful approach to multiple testing[J]. Journal of the Royal Statistical Society, Series B (Methodological), 1995, 57(1): 289-300.

[9] WANG J F, ZHOU H, HAN L Y, et al. Traditional Chinese medicine information database[J]. Clinical Pharmacology & Therapeutics, 2005, 103(3): 501-501.

[10] WHEELER D L, CHURCH D M, FEDERHEN S, et al. Database resources of the National Center for Biotechnology[J]. Nucleic Acids Research, 2003, 31(1): 28-33.

[11] HAMOSH A, SCOTT A F, AMBERGER J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic Acids Research, 2005, 33: D514-517.

[12] AMBERGER J S, BOCCHINI C A, SCHIETTECATTE F, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders[J]. Nucleic Acids Research, 2015, 43: D789-798.

[13] GROZA T, KOHLER S, MOLDENHAUER D, et al. The human phenotype ontology: semantic unification of common and rare disease[J]. The American Journal of Human Genetics, 2015, 97(1): 111-124.

[14] KOHLER S,VASILEVSKY N A,ENGELSTAD M,et al. The human phenotype ontology in 2017[J]. Nucleic Acids Research,2016,45(D1): D865-D876.
[15] KOHLER S,DOELKEN S C,MUNGALL C J,et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data[J]. Nucleic Acids Research,2014,42: D966-974.
[16] ROBINSON P N,MUNDLOS S. The human phenotype ontology[J]. Clinical Genetics,2010,77(6): 525-534.
[17] KANEHISA M,FURUMICHI M,TANABE M,et al. KEGG: new perspectives on genomes,pathways,diseases and drugs[J]. Nucleic Acids Research,2016,45(D1): D353-D361.
[18] 韩增叶,田平芳. KEGG 数据库在生物合成研究中的应用[J].生物技术通报,2011(01): 76-82.
[19] KANEHISA M,SATO Y,FURUMICHI M,et al. New approach for understanding genome variations in KEGG [J]. Nucleic Acids Research,2018,47(D1): D590-D595.
[20] 李向真,刘子朋,李娟,等. KEGG 数据库的进展及其在生物信息学中的应用[J].药物生物技术,2012,19(06): 535-539.
[21] JANET P,ÀLEX B,NURIA Q R,et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants[J]. Nucleic Acids Research. 2017,45(D1): D833-D839.
[22] PINERO J,QUERALT R N,BRAVO À,et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes [J]. Database: the Journal of Biological Databases and Curation, 2015, 2015: bav028.
[23] QUERALT R N,PINERO J,BRAVO À,et al. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases[J]. Bioinformatics,2016,32(14): 2236-2238.
[24] BRAVO À,PINERO J,QUERALT R N,et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research[J]. BMC Bioinformatics,2015,16(1): 55.
[25] Li Y H,Yu C Y,Li X X,et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics[J]. Nucleic Acids Research,2017,46(D1): D1121-D1127.
[26] BERNSTEIn F C,KOETZLE T F,WILLIAMS G J,et al. The Protein Data Bank: A computer-based archival file for macromolecular structures[J]. European Journal of Biochemistry,1977,185(2): 584-591.
[27] SUSSMAN J L,LIN D,JIANG J,et al. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules[J]. Acta Crystallographica Section D Biological Crystallography,1998, 54: 1078-1084.
[28] 王三山. 生物大分子空间结构数据库(PDB)简介[J]. 生命的化学(中国生物化学会通讯),1991(2): 13-15.
[29] BREITKREUTZ B J,STARK C,TYERS M. The GRID: the general repository for interaction datasets [J]. Genome Biol. 2003,4(3): R23.
[30] SAFRAN M,SOLOMON I,SHMUELI O,et al. GeneCards ™2002: towards a complete,object-oriented,human gene compendium[J]. Bioinformatics,2002,18(11): 1542-1543.
[31] SAFRAN M,DALAH I,ALEXANDER J,et al. GeneCards Version 3: the human gene integrator[J]. Database, 2010: baq020-baq020.
[32] STELZER G,ROSEN N,PLASCHKES I,et al. The GeneCards suite: from gene data mining to disease genome sequence analyses[J]. Current Protocols in Bioinformatics,2016,54(1): 1.30. 1-1.30. 33.
[33] RAPPAPORT N,TWIK M,PLASCHKES I,et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search [J]. Nucleic Acids Research, 2017, 45 (D1): D877-D887.
[34] STARK C,BREITKREUTZ B J,REGULY T,et al. BioGRID: a general repository for interaction datasets[J]. Nucleic Acids Research. 2006,34: D535-D539.
[35] 王建.蛋白质相互作用数据库[J].中国生物化学与分子生物学报,2017,33(08): 760-767.
[36] 余鑫煜,许正平. 蛋白质相互作用数据库及其应用[J].中国生物化学与分子生物学报,2008(03): 189-196.
[37] SALWINSKI L,EISENBERG D. Computational methods of analysis of protein-protein interactions[J]. Current

Opinion in Structural Biology,2003,13(3): 377-382.

[38] XENARIOS I,EISENBERG D. Protein interaction databases & search[J]. Protein Technologies and Commercial Enzymes,2010,145: 334-339.

[39] XENARIOS I,RICE D W,SALWINSKI L,et al. DIP: the database of interacting proteins[J]. Nucleic Acids Research,2000,28(1): 289-291.

[40] SALWINSKI L,MILLER C S,SMITH A J,et al. The database of interacting proteins: 2004 update[J]. Nucleic Acids Research,2004,32(suppl_1): D449-D451.

[41] KERRIEN S,ARANDA B,BREUZA L,et al. The IntAct molecular interaction database in 2012[J]. Nucleic Acids Research,2011,40(D1): D841-D846.

[42] HERMJAKOB H,MONTECCHI P L,LEWINGTON C,et al. IntAct: an open source molecular interaction database[J]. Nucleic Acids Research,2004,32(suppl_1): D452-D455.

[43] VON M C,JENSEN L J,SNEL B,et al. STRING: known and predicted protein-protein associations,integrated and transferred across organisms[J]. Nucleic Acids Research,2005,33(suppl_1): D433-D437.

[44] VON M C,JENSEN L J,KUHN M,et al. STRING 7—recent developments in the integration and prediction of protein interactions[J]. Nucleic Acids Research,2006,35(suppl_1): D358-D362.

[45] SZKLARCZYK D,MORRIS J H,COOK H,et al. The STRING database in 2017: quality-controlled protein-protein association networks,made broadly accessible[J]. Nucleic Acids Research,2017,45(D1): D362-D368.

[46] ENSEN L J,KUHN M,STARK M,et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms[J]. Nucleic Acids Research,2008,37(suppl_1): D412-D416.