学校编码：10384　　分类号＿＿＿＿密级 ＿
学号：21620141152535　　UDC ＿

厦门大学

硕 士 学 位 论 文

药物诱导毒性图谱数据库的构建

Construction of Drug-Induced Toxicity related Profile Database

黄丽红

指 导 教 师 姓 名：纪 志 梁 教 授

专 业 名 称：生物化学与分子生物学

论 文 提 交 日 期：2017 年 月

论 文 答 辩 时 间：2017 年 月

学 位 授 予 日 期：2017 年 月

答辩委员会主席：＿＿＿＿＿＿

评 阅 人：＿＿＿＿＿＿

2017 年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（ 纪志梁 教授 ）课题（组）的研究成果，获得（ 纪志梁 教授 ）课题（组）经费或实验室的资助，在（ 厦门大学生物信息学辅助药物开发 ）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

目录

Contents

摘要

药物不良反应（Adverse drug reaction, ADR）是临床医学和公共卫生领域的重大难题之一。它不仅影响临床药物治疗效果，而且对病患的健康造成不同程度的危害甚至是死亡；它同时也是新药研发失败和药物退市的主要因素之一。

由于缺乏足够的基础信息支持，目前对药物不良反应分子机制研究存在较大的局限性。本课题针对 ADR 发生的几种机制，收集了蛋白质-ADR、遗传变异-ADR 和基因调控-ADR 三个层次的关系数据。我们在上一代药物诱导毒性相关蛋白质数据库（Drug-Induced Toxicity Protein）DITOP 的基础上，广泛地从 PubMed 文献中提取报道的蛋白引起的药物不良反应关系信息，建立蛋白质-ADR 关系数据；收集和挖掘 DrugBank、GWAS Catalog、Allele Frequency Net Database 和 Ensembl 四个数据库中的相关信息，获得 ADR-遗传变异关系数据；提取基于美国分子指纹图谱项目（Library of Integrated Cellular Signatures）LINCS 的数据挖掘得到的 ADR-基因调控关联数据。此外，我们还整合了其他可靠资源获取的药物、蛋白质、遗传变异、基因调控和 ADR 多种类型相关数据，构建了新一代药物诱导毒性图谱数据库（Drug-Induced Toxicity Profile）DITOP2。

DITOP2 数据库构建于 Linux 操作系统之上，底层是 Oracle 10g 数据库管理系统。为方便用户访问，我们设计了包括关键词检索和浏览在类的多种数据访问方式。以此同时，我们采用了以数据驱动文档（Data-Driven Documents）D3.js 技术为蓝本的数据可视化技术，以动态网络等方式展示数据。

目前，DITOP2 数据库共收录了标准化 ADR 术语 1107 个，其中一级 SOC（System Organ Classes）词条 11 个，二级 HLGT（High Level Group Terms）词条 17 个，三级 HLT（High Level Terms）词条 749 个，四级 PT（Preferred Terms）词条 556 个。基于蛋白质、基因调控和基因变异的 ADRs 分别有 373、709 和 110 个。数据库收集了药物 1547 个，分布在以上三个水平的数据分别有 434、1,349 和 118 个。非冗余的 ADR 与蛋白质、遗传变异、基因调控关系共 547,173 对，其中分别包括药物-蛋白质-ADR 关系有 2,692 对；药物-遗传变异-ADR 关系有 3,479 对；药物-基因调控-ADR 关系有 541,002 对。初步统计结果表明，药物不良反应

的较易发生于肠道和神经系统。蛋白质介导的药物不良反应主要发生在代谢和神经系统；而由于基因变异所引发的 ADR 集中发生在皮肤和免疫系统。

药物诱导毒性图谱数据库 DITOP2 的建立将会为药物开发或者医疗领域的研究提供多层次，多角度的信息库，辅助 ADR 相关机制的研究和预测，进一步促进精准医疗和个性化用药的发展。

关键词：药物不良反应；数据库；精准医学

Abstract

Adverse drug reaction (ADR) is a severe medical and social problem that needs to be solved immediately. Every year, it leads to not only substantial pains to patients but also a number of new drug discovery failures.

One of the key factors that slows down ADR mechanism study and assessment is the shortness of sufficient supporting information for ADRs. Therefore, in this study, we collect three levels of molecular information on purpose for understanding ADR mechanism, including protein, gene and genetic variation. For instance, we extract protein/gene-ADR associations from various public resources like PubMed and DrugBank, genetic variation-ADR association from GWAS Catalog and literature, gene regulation-ADR associations from CTD database and data mining of transcriptomes in LINCS (Library of Integrated Cellular Signatures).

Upon these data, we make a significant upgrade of the Drug induced Toxicity related Protein database (DITOP) to its version 2. DITOP2 collects 1,107 standard ADR terms, including 11 SOC (System Organ Classes) terms, 17 HLGTs (High Level Group Terms), 749 HLTs (High Level Terms), and 556 PTs (Preferred Terms). Of these ADRs, 373, 709 and 110 ADRs show associations with protein, gene and genetic variation, respectively. Also, of 1,547 drugs covered in this database, 434, 1,349 和 118 drugs likely induce ADRs via protein, gene and genetic variation, respectively. Totally, DITOPs deposit 547,173 association pairs, including 2,692 drug-protein-ADR relations, 3,479 drug-genetic variation-ADR relations, and 541,002 potential drug-gene-ADR relations.

DITOP2 is running on Linux operation system, managing by the Oracle 10g DBMS. We design the user interface using the Java script language and visualize the data by D3.js (Data-Driven Documents) technology. The database supports both key search and browse method for data retrieval. Batch data download is also allowed.

Furthermore, we also carry out preliminary analysis on DITOPs. It looks that protein-mediated ADRs often happen in metabolism and nerve system; however, the genetic variation-mediated ADRs often occur in skin and immune system.

In summary, DITOP2 will aid both ADR mechanism study and drug discovery. It will be of value to future precision medicine, especially in refining drugs with better ADR profile.

Keywords: Adverse drug reaction; Database; Precision medicine

第一章 前言

1.1 药物不良反应

药物的一般定义是在人类或动物中，用于诊断，缓解，治疗或者预防疾病的任何物质。 药物的使用本身具有两面性，药物在进入机体内，往往不能特异性地作用于患部以及靶蛋白，所以药物在发挥药效的同时，常常伴随一些不希望出现的有害反应，通常称之为药物不良反应。药物不良反应是导致危害和死亡比较常见的一种因素。根据世界卫生组织，WHO（World Health Organization）国际药物监测中心定义，药物不良反应（Adverse Drug Reactions，ADRs）是指正常剂量的药物用于预防、诊断、治疗疾病或调节生理机能时出现的非预期有害反应[1]。药物不良反应一般可以分为基本的两大类[2]：A 类型是一种靶向的反应，这种反应一般可以预测，与常规的药理作用有关，有一个比较明显的剂量反应关系，在停药或减量后症状很快减轻或消失，死亡率较低。其主要表现有过度作用，毒性、首剂效应、继发、停药综合症、后遗效应；B 类型则是种脱靶的反应，该类型很难预测，通过常规的毒理学筛选不易发现明确的剂量反应关系，一般在药物被批准上市之后才能监测到, 更为详细的分类是由 WHO 发展出的六类分类法[2]，如表 1.1 所示。

表 1.1 六类分类法

Table 1.1 Edwards and Aronson classification of ADRs

ADR 分类	特性	例子
A 类(Augmented)	剂量相关	三环类抗抑郁药的抗胆碱反应
B 类(Bizarre）	剂量不相关	青霉素过敏反应
C 类(Chronic）	剂量累积相关	糖皮质激素引起下丘脑-垂体-肾上腺轴抑郁
D 类(Delayed)	延迟性	己烯雌酚引起阴道腺癌
E 类(Endofuse)	停药性	吗啡戒断综合征
F 类(Failure)	治疗失败	口服避孕药失败

现实生活中，药物不良反应的发生率相当高，其所带来的影响不只是药效的

降低，还危害到人体健康及至生命。在人类的医药史上，多次发生的药物不良反应的危害程度震惊全球，有导致数千名婴儿畸形和死亡的反应停（沙利度胺，Thalidomide）药害事件；震惊国人的中药马兜铃酸所致的多人罹患晚期肾衰竭事件等。据相关部门的统计，美国每年因 ADR 事件住院的人数超过 200 万人，死亡人数超过 10 万，占社会人口死亡原因的第四至第六位[3, 4]，这种情况还在持续恶化。就美国食品药品监督管理局（FDA）官方统计的 ADR 自愿报告数据，严重不良反应事件包括死亡人数，在过去的 10 年期间增加了三倍[5]。相较于美国，西方各国由于药物不良反应事件致在死的人数也是有增无减[6]。据有关文献报道，中国每年因为药害事件住院人数约有 250 万[7]，死亡人数有 20 万，这样大大小小的药害事件不胜枚举。

药物不良反应发生因素有如下：药物本身因素，包括药物有效成分的代谢产物、增溶剂、稳定剂、添加剂、赋型剂、着色剂、合成中产生的杂质等，均可导致药物不良反应；药代动力学因素，药物进入人体需要通过吸收，分布，代谢，排泄等过程，任何一个环节都有可能产生不良反应；个体因素，主要是与个体之间的特异性遗传因素有关，如图 1.1 所示，该图是整理了 FDA 上的数据所得的。

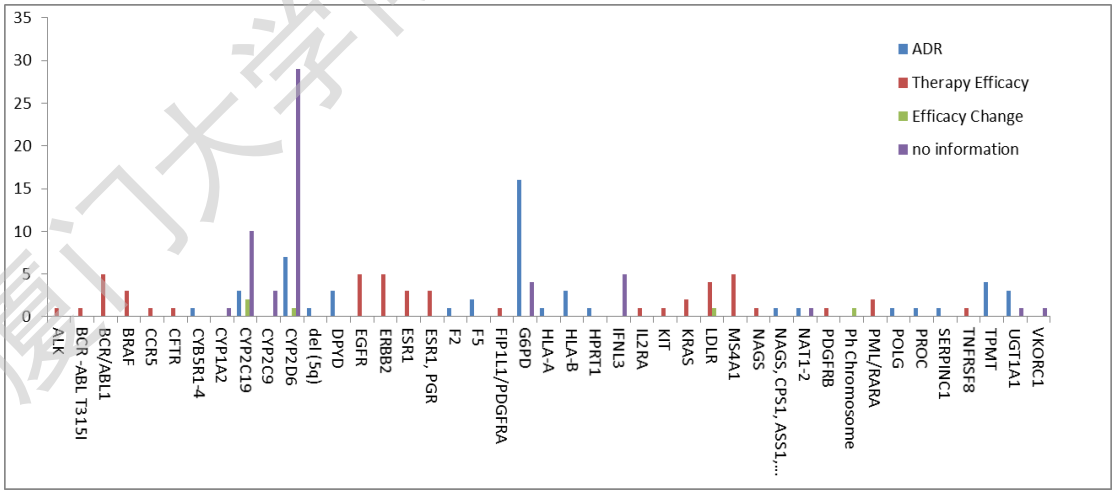

图 1.1 遗传标识类型及对应的药物反应

Fig 1.1 Gene genotype and drug response

美国 FDA 药品药单上的遗传标识一个很重要的方面即遗传变异如何导致 ADR 反应。绝大多数原因是由于酶活性的改变，包括失活、减弱或者增强导致：酶活性缺失或减弱的情况有，CYP450 蛋白家族的对药物代谢能力较弱可能引起的 ADR，葡萄糖-6-磷酸脱氢酶（G6PD）失活，导致个体在服用药物后引起的溶血性贫血，巯嘌呤甲基转移酶（TPMT）活性丧失或减弱导致的骨髓毒性，二氢嘧啶脱氢酶（DPD）活性缺失导致分解代谢减弱引起严重的 ADR，尿苷二磷酸葡糖醛酸转移酶（UGT1A1）活性缺失或减弱个体在服用药物后胆红素的排出减少引起的高胆红素血症，NADH-细胞色素还原酶 CYB5R1-4 缺失引起的高铁血红蛋白血症，凝血因子 F5 Leiden 和凝血酶原 F2 突变 G20210A 携带者以及抗凝血酶 3 缺失，患者在服用某些药物的时候导致的凝血反应，次黄嘌呤鸟嘌呤磷酸核糖基转移酶（HGPRT）缺失导致的痛风和肾脏疾病，尿素循环相关酶缺失导致的高血氨脑病甚至死亡，N-乙酰转移酶酶活性减弱可能导致的毒性反应，线粒体 DNA 聚合酶（POLG）突变引起的急性肝功能衰竭最终导致死亡；酶活性增强的情况有 CYP2D6 基因多拷贝引起的 Codeine 过快地转化为 Morphine 而导致的儿童呼吸抑制甚至死亡。另一方面，I 型人类白细胞抗原如 HLA-B*5701、HLA-B*1502、HLA-A*3101 等位基因的携带者服用某些药物将导致严重的过敏反应。

对于药品不良反应的监测系统，早在 1968 年，WHO 就制订了一项国际药物监测合作试验计划，该计划有 10 个国家参与收集和交流 ADRs 报告，制定 ADRs 报表和术语，药品目录，形成计算机报告管理系统。而后 1970 年设立，1997 年更名为乌普沙拉监测中心（Uppsala Monitoring Centre，UMC）也对了解和评估药物安全提供了重要的依据[8]。一些其他国家也纷纷建立起自己监测管理系统，如美国疫苗不良事件报告系统（Vaccine Adverse Event Reporting System，VAERS）[9] 和美国 FDA 不良事件报告系统（FDA Adverse Event Reporting System，FAERS）[10] 负责收集、分析、管理上市后药物不良事件；欧盟成立的欧洲专利药品委员会（Committee for Proprietary Medicinal Products, CPMP），其承担欧共体药物警戒的职责，而后在 1993 年又成立欧洲药品评价局（European Medicines Evaluation Agency, EMEA）；包括澳大利亚的蓝卡系统、英国的黄卡系统[11]、日

本的药品安全数据报告系统等。而在我国 ADRs 监测工作起步较晚，到 2002 年底，全国范围的药品不良反应监测体系初步形成。目前，国际上面对 ADRs 的监测方法有很多，都旨在及时准确的报告 ADRs。传统的 ADRs 监测方法包括医院集中监测、病例对照研究、处方事件监测、自发呈报、医学记录应用等。随着技术的发展，计算机广泛地应用于药物不良反应监测，相继出现一些自动检测软件，药物不良事件的 报告数量远多于单纯只由医生、药师、护士自动呈报的药物不良反应事件。

1.2 精准医学的发展

2015 年 1 月 20 日，美国总统奥巴马在国情咨文中提出了“精准医学（Precision Medicine）”计划，而后美国的 FDA 出台了 3 个草案，从基于二代测序的遗传性疾病的体外诊断、公众数据库的使用及药物/诊断共同开发等方面支持它。随着基因测序技术与大数据分析技术的迅速发展，各国也纷纷开展自己的精准医学计划。2015 年 3 月中国科技部即提出中国精准医学计划，并正式写入十三五规划。据相关资料显示，截止到目前，中国精准医疗的公司已有 854 家，这还不包括正在研究所及大学的实验室中酝酿的项目。精准医疗产业链包括液体活检、基因编辑、大数据、以及免疫细胞治疗等。英国宣布启动针对癌症还有罕见病患者的英国十万人基因组计划，旨在推进基因组医疗整合至该国医疗服务体系（NHS），从而提升有利于病患的诊断及精准治疗；法国宣布投资 6.7 亿欧元来启动基因组和个体化医疗项目；2016 年 5 月澳大利亚则宣布了零儿童癌症计划（Zero Childhood Cancer Initiative）。韩国政府则在 2015 年 11 月宣布启动万人基因组计划。如今，精准医学已经是全球关注的事业，归根到底这还是基因和大数据的发展的必然结果。

所谓精准医疗，本质上是一种个性化医疗，是指根据人体基因组信息为基础，结合多组学包括代谢组、蛋白质组等相关内环境信息和环境以及生活习惯进行疾病治疗和干预的最佳方式。美国总统奥巴马之前这样解释精准医疗：“把按基因匹配癌症疗法变得像匹配血型那样标准化，把找出正确的用药剂量变得像测量体温那样简单，总之，每次都给恰当的人在合适的时间使用合适的疗法，达到治疗

效果最大化，医疗资源最优化，损害作用最小化。而损害最小化的重要内容就是减少药物不良反应的发生，将潜在的不良反应最小化，达到更加精确的诊断，提供最有效的治疗。

虽然精准医疗的前景很好，但是实现过程却极其复杂。美国国立卫生研究院（NIH）主任弗朗西斯•柯林斯曾经说过，该计划需要整合人类多组学（基因组、蛋白质组、代谢组等等）及高通量测序技术、计算生物学分析、临床和医学信息学、疾病的特异性动态标志物与网络、毒性敏感监测、精准药物的研发、药物疗效依赖性治疗还有预测预后，从而更加精准改善个体健康。同时他还特别强调，采集各种信息、整合多种数据是一项首要的工作，而且处理这些信息的工具必须是简单易用的，数据信息必须是准确的，他自认为“要实现这两点并不易。个性化用药作为精准医学重要的一部分，研究药物发挥疗效和产生毒副作用的机制就尤其重要，为了实现这样的目标，前提之一是要对药物可能诱发不良反应信息充分的采集，从不同的角度去研究药物不良反应发生的机制。

1.3 药物基因组学研究现状

药物基因组学（Pharmacogenomics，PGx）是精准医学中备受关注的研究领域，在临床用药指导中，很大程度上改善数百万患者的药物治疗效果，在精准医学发展计划中担当了重要的角色。药物基因组学概念的提出将近有数十年时间，早在 20 世纪 90 年代，药物基因组学这一术语就开始在一些科学著作中出现。这源于全基因组学技术的出现与发展。美国药学科学家协会（AAPS）对 PGx 的定义是“全基因组水平分析药物效应和毒性的遗传标记”[12]。该定义阐明药物作用包括疗效和毒副作用及药物代谢、转运和药物靶点的基因多态性之间的关系。

药物基因组学主要是建立在遗传学基础上的研究，遗传的主要物质是 DNA，而 DNA 是携带个体遗传基因和传递遗传信息最基本的物质。自然界中，同一种生物常会在某些方面有所差异，导致多态性（Polymorphism）现象的出现。多态性的出现是多个不同等位基因作用导致的结果，而遗传多态性则是基于变异频率超过 1%的情况。目前，多态性的可能存在形式有几种，由单个核苷酸碱基（A、T、C、G）发生的变异定义为单核苷酸多态性（single nucleotide polymorphism ，