PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2011, 38(1): 11~19
www.pibb.ac.cn

# 基于生物信息学方法发现潜在药物靶标 *

刘 伟 谢红卫 **
(国防科学技术大学机电工程与自动化学院自动控制系，长沙 410073)

**摘要** 药物靶点通常是在代谢或信号通路中与特定疾病或病理状态有关的关键分子．通过绑定到特定活动区域抑制这个关键分子进行药物设计．确定特定疾病有关的靶标分子是现代新药开发的基础．在药物靶标发现的过程中，生物信息学方法发挥了不可替代的重要的作用，尤其适用于大规模多组学数据的分析．目前，已涌现了许多与疾病相关的数据库资源，基于生物网络特征、多基因芯片、蛋白质组、代谢组数据等建立了多种生物信息学方法发现潜在的药物靶标，并预测靶标可药性和药物副作用．

**关键词** 药物靶标，网络特征，基因芯片，蛋白质组学
**学科分类号** Q61
**DOI**: 10.3724/SP.J.1206.2010.00251

药物靶标是指体内具有药效功能并能被药物作用的生物大分子，如某些蛋白质和核酸等．靶标基础上的药物开发流程将组织作为一系列基因和通路的集合，目标是发展一种能够影响一个基因或者分子机制(即一个靶标)的药物，治疗疾病引起的缺陷同时尽可能地减少副作用．尽管实验技术取得了很大进步，人们对于生物系统有了更深入的理解，药物发现仍旧是个漫长的过程，新药研发昂贵、困难并且低效．其中，药靶发现是非常重要的一个限速步骤．

药靶筛选和功能研究是发现特异的高效、低毒性药物的前提．靶标发现与确证的一般流程(图 1)是：利用基因组学、蛋白质组学以及生物芯片技术等获取疾病相关的生物分子信息，并进行生物信息学分析，然后对相关的生物分子进行功能研究，以确定候选药物作用靶标，针对候选药物作用靶标，设计小分子化合物，在分子、细胞和整体动物水平上进行药理学研究，验证靶标的有效性．

常见的用于药靶发现的实验方法包括：微生物基因组学、差异蛋白质组学、核磁共振(NMR)技术、细胞芯片技术、RNAi 技术、基因转染技术和基因敲除动物等．随着组学数据的积累，仅凭实验方法已经不能满足高通量大规模数据分析的需求．在药物研发过程中，生物信息学方法对于相关数据

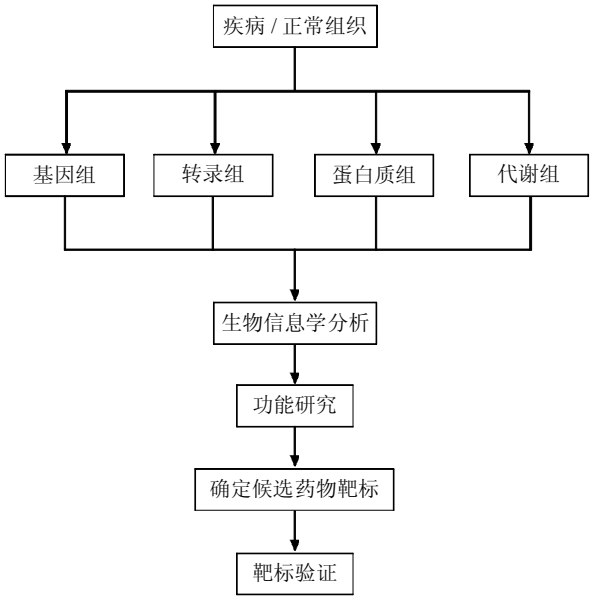


**Fig. 1 Typical flow chart of drug target discovery**
**图 1 药物靶标发现的典型流程**

* 国家自然科学基金(60773021, 60603054, 30800200, 30621063), 国家重点基础研究发展计划(973)(2006CB910803, 2006CB910706), 国家高技术研究发展计划(863)(2006AA02A312), 国家科技重大专项(2008ZX10002-016, 2009ZX09301-002)和蛋白质组学国家重点实验室课题(SKLP-Y200811)资助项目.
** 通讯联系人.
Tel: 0731-84573369, E-mail: xhwei65@nudt.edu.cn
收稿日期：2010-05-10，接受日期：2010-08-18

的存储、分析和处理，以及如何有效地发现和验证新的药靶，发挥了重要的作用．本文首先介绍可用于药靶发现的数据库资源，包括疾病相关的基因数据库、候选药靶数据库和基因芯片数据库等，其次讨论了基于多种组学数据进行药物靶标发现的生物信息学方法，如基于基因组、基因表达谱、蛋白质组、代谢组的方法以及整合多组学数据的系统生物学方法，再次描述了生物信息学方法在药物靶标验证方面的应用，主要是预测蛋白质可药性以及药物副作用，最后是总结和展望．

## 1 用于药靶发现的数据库资源

### 1.1 疾病相关的基因数据库

当研究某个基因时，人们最感兴趣的问题之一是: 它是否与疾病相关? 有两种方法可以查询这个信息，一是通过数据库查询基因与疾病的相关性．二是，如果该基因与疾病的关系未知，可以尝试将基因在染色体上的位置与疾病进行对应．目前，已有一些数据库存储了与疾病相关的基因信息，方便研究人员对相关的基因或蛋白质进行查询和比较．

与人类疾病相关的基因以及基因敲除时的异常情况存储在 OMIM(Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim/)、LocusLink 和 The Human Gene Mutation 等数据库中．其中，OMIM 是分子遗传学领域最重要的生物信息学数据库之一．该数据库是人类基因和遗传性疾病的电子目录，提供疾病与基因、文献、序列记录、染色体定位及相关数据库的链接．该数据库可以通过 ENTREZ 进行搜索，并且利用“limit”选项限制所搜索的染色体或类别等．

其他与疾病相关的基因数据库还有 COSMIC (www.sanger.ac.uk/genetics/CGP/cosmic)、Cancer Gene Census(www.sanger.ac.uk/genetics/CGP/Census)等．COSMIC 数据库存储了癌症相关的候选基因，提供体内基因敲除信息以及人类癌症的相关细节．Cancer Gene Census 项目对癌症相关的基因进行分类，这些基因在敲除时与癌症表现出可能的因果关联．而 GeneRif 系统提供与疾病高度相关基因的注释信息[1]．此外，基因组规模的关联数据库、遗传关联数据库和小鼠基因敲除数据库等也为基因查询提供了丰富的注释信息．

### 1.2 候选药靶数据库

相比疾病相关的基因，已知药物靶标的数目要少的多．通过对已成功应用药物的靶标进行鉴别，TTD(Therapeutic Target Database)数据库提供已知的诊疗目标、疾病条件和对应的药物．DrugBank (www.drugbank.ca)作为一个有用的生物信息资源，结合了详细的药物数据和综合的药物靶标信息，提供美国食品与药物协会的研究中正在进行测试的药物和对应的靶标．PDTD (Potential Drug Target Database)通过文献和数据库挖掘的方式，收集了超过 830 个已知或潜在的药物靶标，并提供蛋白质结构、相关疾病和生物学功能等信息[2].

### 1.3 疾病相关的基因芯片数据库

基因芯片数据库是药物靶标发现的重要来源，人们已经建立了一些专门的数据库用于存储疾病相关的基因芯片数据．GEO(Gene Expression Omnibus)作为存储基因芯片的主要数据库资源，包含了丰富的癌症相关的基因芯片数据．当查询“Homo sapiens”和“Cancer”时，返回了 278 个数据集．2003 年 10 月，Daniel 等建立了 ONCOMINE 数据库(http://www.oncomine.org)，专门收集癌症相关的基因芯片数据集，提供在网页基础上的数据挖掘和基因组规模的表达分析．在 ONCOMINE 3 版本中，该数据库包含了 264 个基因表达数据集，超过 2 万个癌症组织和正常组织的样本数据[3]．其他基因芯片数据库包括斯坦福基因芯片数据库 (http://genome-www5.stanford.edu/MicroArray/SMD)、EBI 芯片表达数据库(http://www.ebi.ac.uk/arrayexpress)，以及 MIT 癌症基因组工程(http://www.broad.mit.edu/cancer/)等，都是药靶发现的重要资源．

### 1.4 其他相关数据库

药物靶标通常具有特定的生物学功能，分析基因的分子类型(例如酶)、亚细胞定位(例如细胞表面)和生物学通路(例如血管新生)对于预测潜在药靶具有重要意义．基因本体论(GO，Gene Ontology，http://www.geneontology.org)和京都基因与基因组百科全书数据库(KEGG，Kyoto Encyclopedia of Genes and Genomes Pathways, http://www.genome.ad.jp/kegg)提供了多个物种中基因的生物学功能、定位和通路信息．同时，有关蛋白质相互作用网络和生物学通路的数据库资源非常丰富，如 DIP、Reactome、NCI (Nature Pathway Interaction Database)、HPRD 和 Biotarca 等，更多的数据库列表可以参考 http://www.pathguide.org．此外，有些数据库专门存储生物学网络的定量数据，例如 BioModels[4]和 JWS online[5]数据库收集了各种化学反应网络的数学模型，并且规模一直在稳步增加．

## 2 用于药靶发现的生物信息学方法

### 2.1 基因组方法

丰富的基因组学数据为药靶发现提供了基础，目前已有多种方法可用于寻找新的药物靶标[6]. 其中，最常用的方法是同源搜索，采用序列比对软件寻找候选基因与已知癌症基因之间的序列同源性，如 BLAST 或基于隐马尔科夫的 HMMER 软件包等. 然而，新的靶标与已知癌症基因的序列可能并不相似. 因此，有必要分析已知药靶中更为普遍的结构特征，如信号肽、跨膜结构域或蛋白激酶域. 此类生物信息学工具包括预测信号肽的 SignalP 和预测跨膜结构域的 TMHMM. 此外，还可以使用基因预测程序从人类基因组序列中预测新基因，寻找全新的药物靶标，常用的程序是 Genescan 和 Grail.

通过单基因敲除实验能够发现生物体中的必要基因(essential gene). 但以必要基因作为癌症治疗的靶标不仅能杀死癌细胞，对于健康细胞也可能是致命的. 因此，大多数以单基因作为靶标的药物治疗是失败的. 双基因的合成致死性(synthetic lethal)为抗癌药物的研究提供了新的前景. 给定一个癌症相关的基因，如果该基因在癌细胞中功能缺失或者功能降低，那么以它的合成致死对象作为药靶就能构成肿瘤细胞的致死条件，同时降低对健康细胞的损伤. 目前，仅在酵母中通过大规模的实验建立了全基因组的合成致死网络. 通过同源预测等方法，Conde-Pueyo 等[7]重建了人的基因合成致死网络，为抗癌研究中候选基因靶标的筛选提供依据.

目前已知的单基因病种类较少，仅限于基因组方法得到的药物靶标作用效果往往不够理想. 随着后基因组时代的到来，其他组学数据在药物靶标发现中发挥了越来越重要的作用.

### 2.2 基因芯片方法

基因芯片技术指将大量(通常每平方厘米点阵密度高于 400)探针分子固定于支持物上与标记的样品分子进行杂交，检测每个探针分子的杂交信号强度，进而获取样品分子的数量和序列信息. 由于基因芯片技术的高通量、快速、平行化等特点，使得疾病相关的基因芯片数据资源非常丰富，利用基因芯片数据挖掘潜在药物靶标成为一种重要的途径. 例如，在 GEO 数据库的基础上，Hu 等[8]建立了大规模的疾病 - 药物对应网络，帮助有效地识别药物靶标.

但由于基因芯片本身存在重复性较差和数据质量不高等问题，需要发展多种有效的分析方法，尤其是能够处理多个数据集、对噪声不敏感的统计方法，以提取海量数据中蕴含的有用信息.

**2.2.1** 基于比较基因芯片数据.

基因芯片能够一次性地记录疾病状态下成千上万个基因的变化情况. 通过比较疾病组与正常组的基因芯片数据，寻找显著差异的基因集合，可用于预测相关的生物标志物或药物靶标. 其中，寻找差异表达基因的计算方法很多，最直接的方法是测量变化倍数，即计算两个样本之间同一个基因的表达量之比. 尽管变化倍数方法直观有效，但是该方法没有考虑噪声和生物学可变性，尤其是癌症这种本质上多相异质的复杂疾病. 因此，更加通用的办法是采用尽可能多的疾病样本进行统计学分析，如 ANOVA 和 T-like 检验等.

进一步，由于单个基因难以检测疾病状态下翻译模型的变化，生物标志物通常包括一组基因，需要一定的聚类方法寻找相关基因的组合. 如 GSEA (Gene Set Enrichment Analysis)方法能够评估两种生物学状态下一组基因集合的统计显著性，已广泛地应用于基因芯片数据的分析[9].

**2.2.2** 多种来源的基因芯片数据的整合.

由于单个芯片数据本身存在的噪声及系统偏差，预测结果往往存在误差. 因此，最新的研究通过整合不同实验来源的多组基因芯片的数据，减少单个芯片实验中的误差影响，寻找更加通用的生物标志物和药物靶标[10-13].

数据整合的目的是将不同来源的芯片数据进行处理，使得相同基因的数据可以相互比较. 在预处理过程中，不同的标准化方法会影响不同来源的芯片数据之间的可比性. Autio 等[10]比较了来自于 5 个芯片组的 6 926 个基因表达数据，评估 5 种标准化方法的应用效果. 经过研究发现，采用 AGC 方法(array generation based gene centering normalization)先进行样本内标准化再进行样本间的标准化时，能够得到最好的预处理结果，即在数千个样本之间得到可比较的基因表达量. 此外，Stafford 等[11]从以下 3 个方面对 8 种常用的标准化方法进行比较: 敏感性和通用性、功能 / 生物学解释以及特征选择和分类错误，方便用户挑选合适的标准化方法进行跨实验室、跨平台的基因芯片表达数据的比较.

采用一定的统计方法对不同来源的芯片数据进行整合，可以在进行更少实验的情况下更好地利用

已有芯片数据，有助于发现多种癌症样本中共同的生物标志物以及某种癌症特异的生物标志物．其中，最简单的方法是 Z 打分归一化．较复杂的方法是提取不同数据集中表达数据的分布特征参数，根据这些特定的参数进行数据集匹配，包括: Distance Weighted Discrimination、Combatting Batch effects、disTran、Median Rank Score、Quantile Discretizing 和 Z 打分变换等．其中，经典方法的是 Daniel 等最早提出的荟萃分析(Meta-analysis)方法[12]．利用 ONCOMINE 数据库，他们收集了 40 个独立数据集(超过 3 700 个芯片实验)，提出了一种独立于单个数据集的统计量 *Q*-value，寻找多种来源数据集中显著差异表达的基因作为荟萃标志物(Meta-signature)．此后，多基因芯片融合方法得到了普遍关注，各种统计方法被用于发现通用标志物并与 Meta-analysis 方法进行比较．例如，Xu 等[13]收集和整合了 26 个公开发表的癌症数据集，包括 21 个主要的人类癌症类型的 1 500 个基因芯片数据，应用 TSPG(Top-Scoring Pair of Groups)分类器和重复随机采样策略，识别通用的癌症标志物．评估结果表明，采用一定的统计方法整合多种芯片数据能够识别出更加稳健的癌症标志物，相比单基因芯片得到的标志物，其将癌症类型与正常组织的区分效果更好．

### 2.3 蛋白质组学方法

通常，功能蛋白的表达异常和调节异常是癌症发生的分子标志，这些决定个体生物性状、代谢特征和病理状况的特殊功能蛋白可以作为潜在的药物靶标．尽管 90%的已知药靶为蛋白质，但由于数据和技术上的原因，蛋白质水平的药物靶标并不如基因、转录水平的研究广泛．近年来，随着更多蛋白质详细数据的获得，在蛋白质水平上进行药物靶标的开发和验证成为研究的热点．

**2.3.1** 基于蛋白质的理化特性．

在蛋白质的理化属性、序列特征和结构特征上，药靶分子和非药靶分子存在着显著的差异．Bakheet 等[14]的工作具有一定的代表性．他们系统分析了 148 个人类药靶蛋白质和 3 573 个非药靶蛋白质的特性，寻找两者的区别并预测新的潜在药物靶标．人类药物靶标蛋白可以归纳为 8 个主要属性：高疏水性、长度较长、包含信号肽结构域、不含 PEST 结构域、具有超过 2 个 N- 糖基化的氨基酸、不超过一个 O- 糖基化的丝氨酸、低等电点和定位在膜上．以这些特征作为支持向量机的输入，可以在药靶和非药靶类之间达到 96%的分类准确率，并识别出 668 个具有类似靶标属性的蛋白质．

基于蛋白质的理化特性进行药物靶标预测，有利于发现药物靶标的一般特征，方法直接、简单．但该方法受已知药靶的影响较大，在确认药靶的有效性时还需要引入更多的证据支持．

**2.3.2** 基于蛋白质相互作用的网络特征．

癌基因(oncogene)是人类或其他动物细胞(以及致癌病毒)固有的一类基因，又称转化基因，它们一旦活化便能促使人或动物的正常细胞发生癌变．通常，癌基因作为网络的 hub 蛋白参与多种细胞进程，在信号通路中间成为信息交换的焦点．发现新的癌症相关基因是癌症研究的主要目标之一，也是发现潜在药靶的基础．人类基因组规模的蛋白质相互作用数据的快速积累为研究癌基因在细胞网络中的拓扑属性提供了条件．

在蛋白质相互作用网络的基础上，Xu 等[15]提取了节点的 5 个网络特征，包括连接度、1N 指数、2N 指数、与致病基因的平均距离以及正拓扑相关系数(positive topology coefficient)，采用 KNN 方法比较疾病相关基因和对照基因在网络特征上的区别．研究结果证实: 疾病相关基因具有更高的连接度，更倾向与其他的致病基因发生相互作用，而且致病基因之间的平均距离明显低于非致病基因．Ostlund 等[16]通过筛选与已知癌基因高度连接的基因，得到了一个由 1 891 个基因组成的集合．通过交叉验证、分析功能注释偏性和癌症组织中的表达差异进行方法验证，提供了一个较为可信的癌症相关的候选基因列表．该基因列表的规模是已知癌基因数目的 2 倍以上，对于生物标志物和药靶发现具有一定的提示作用．进一步，Li 等[17]通过整合多种数据源识别癌基因，包括网络特征、蛋白质的结构域组成和功能注释信息等．这些研究表明: 根据蛋白质在相互作用网络中的特征，能有效地提示大量的潜在药物靶标，并且方便与其他方法相结合．同时，蛋白质复合物的拓扑属性和模块性也可用于药靶筛选．不同于一般的二元蛋白质相互作用，复合物更接近于细胞内的真实状态．在复合物内部，多肽之间相互连接成为不同的核，其他蛋白质与核发生相互作用形成各种模块．

蛋白质相互作用网络体现了蛋白质组的系统水平描述，对于建模复杂的生物系统具有非常重要的作用．有关蛋白质相互作用的知识可以使人们在分子水平上更好地理解信号转导的生理学活动，以及

由于通路的交叠部分异常造成的多种疾病.

**2.3.3** 比较蛋白质组方法.

蛋白质组学是研究特定时空条件下细胞、组织等所含蛋白质表达谱的有效手段，也是寻找癌症分子标记和药物靶标的重要方法. 相关的蛋白质组学技术包括免疫亲和纯化(affinity purification)、蛋白质活性表达谱(activity-based profiling)和蛋白质芯片(microarray)等，识别与某一特定疾病或者病理条件相关的蛋白质.

基于蛋白质组学研究药靶通常采用比较蛋白质组分析方法，例如稳定同位素差异标记、ICAT (Isotope-coded affinity tag)或 iTRAQ 技术，能够较为精确地定量蛋白质丰度的变化. 通过比较癌症人群与正常人群在对应病理组织 / 器官内蛋白质的差别，挖掘潜在的药物靶标. 例如，Hu 等[18]采用二维液相色谱串联质谱法(2D-LC-MS/MS)比较肺癌患者与正常人的血清蛋白差异，经过蛋白质鉴定和定量分析，发现了 2 078 个蛋白质可能存在差异，进而挑选出 Tenascin-XB(TNXB)作为候选的生物标志物用于预测肺癌的早期转移. 此外，如果不能直接找到对应的活性小分子，也可以通过比较疾病样本和正常样本中蛋白质的表达差异，鉴别发生异常的生物学通路[19]. 采用总体的蛋白质谱方法(如 MudPIT)获取充足的信息，发现与特定表型相关的蛋白质和通路. 定位到相应的生物学通路之后，再从中确定药物靶标.

随着人类蛋白质组计划的推进，蛋白质组技术的发展为系统地、规模化地寻找蛋白质药靶和蛋白质药物提供了有力的武器. 但由于现有数据的规模和质量问题，以及分析方法的限制，采用蛋白质组学方法发现的药物靶标还没有人们预想的多，还有着广阔的发展空间.

## 2.4 代谢组方法

代谢组学是生物体内小分子代谢物的总和，所有对生物体的影响均可反映在代谢组水平. 代谢组放大了蛋白质组的变化，更接近于组织的表型. 代谢途径的异常变化反映了生命活动的异常，因此定量描述生物体内代谢物动态的多参数变化可揭示疾病的发病机制. 通常，代谢组学的实验技术包括核磁共振、质谱、色谱等，其中核磁共振技术是最主要的分析工具，其次是液相色谱 - 质谱联用(LC/MS)和气相色谱 - 质谱联用(GC/MS). 通过 GC/MS 技术解析出代谢物的质谱图，将其与现有数据库进行比较，可以鉴定该代谢化合物. 由于缺少标准的代谢物数据库，该方法的鉴定结果有限. 采用生物信息学方法对代谢组数据进行分析和处理，比较正常组和模型组的区别，可以帮助药靶发现以及药效评估. 如 Pohjanen 等[20]提出了一种名为统计多变量代谢谱(staistical multivariate metabolite profiling)的策略，在代谢 GC/MS 数据的基础上辅助药靶模式发现和机制解释.

同时，代谢组学对于生物标志物发现、药物作用模式和药物毒性研究具有重要作用. 在酶网络的基础上，Sridhar 等[21]发展了一种分支定界(branch-and-bound)方法，命名为 OPMET，寻找优化的酶组合(即药物靶标)，用于抑制给定的目标化合物并减少副作用. 类似的，通过提取代谢系统的特征，Li 等[22]采用整数线性规划模型在整个代谢网络范围内寻找能够阻止目标化合物合成的酶集合，并尽可能地消除对非目标化合物的影响.

## 2.5 整合多组学数据的系统生物学方法

系统生物学将基因组、蛋白质组和代谢组等不同组学的数据进行整合，研究在基因、mRNA、蛋白质、生物小分子水平上系统的生物学功能和作用机制. 对于疾病的发生和发展提供了更好的理解，同时有助于识别药物的作用和毒性、模拟药物作用的过程、发现特异的药物作用靶标.

**2.5.1** 文本挖掘方法.

由于人类疾病背后的生物机制相当复杂，在药靶发现中最重要的任务不仅是要挑选和优化可靠的作用靶点，而且要理解在疾病表型下隐含的分子相互作用，提供可预测的模型并建立人类疾病的生物网络. 因此，需要广泛地收集和过滤现有的各个层面的异质数据和信息. 目前，最流行的生物医学文献数据库 MEDLINE/PubMed 收录了从 1970 年开始的超过 1 800 万篇文献的摘要，并且每月还会新增超过 6 万篇的摘要. 据估计，存储化学、基因组、蛋白质组和代谢组数据的数据库规模每两年就会翻一倍. 如此丰富的生物数据和信息为药靶发现提供了巨大的新机遇.

尽管分子生物学和医学研究中数据库的重要性日益增长，绝大部分的科学论文并非存在于结构化的数据库条目中. 这些知识必然无法为计算机程序所理解，甚至对于人来说都是难以发现的. 文本挖掘方法是机器学习和自然语言处理方面的计算方法，能够有效地用于数据挖掘和知识理解，从海量的医学文献中挖掘与药靶发现相关的有用知识[23]. 其主要内容包括: 识别生物学实体，包括基因、基

因产物、通路和疾病；提取蛋白质相互作用关系，并以网络图形化表示；抽提出特定细胞类型中相关的生物学通路，以及计算机仿真所需的动力学参数；建立存储这些抽提信息的数据库. 目前，生物知识的文本挖掘方法主要采用实体的共出现分析和自然语言处理，已成功地用于疾病相关的网络重建以及生物数据分析，常用软件包括 Protein Corral 和 EBIMed. 进一步，更复杂的文本挖掘方法可以从文献中抽提详细的相互作用注释信息，如 Wang 等[24]发展了一种 CMW(Correlated Method-Word)模型从文本中提取蛋白质相互作用的检测信息.

**2.5.2** 通路建模与仿真.

药物作用是一个复杂的动态过程，如果不能找到合适的方法就很难确认药物的有效性. 例如，在药物开发过程中常用的手段之一是基因敲除实验，其作用方式与在特定酶上的竞争抑制过程完全不同. 在基因敲除过程中，给定的通路可能被完全关闭，也可能由于系统的自身补偿作用而只有部分的影响. 在此基础上设计的靶向药物可能存在效率较低的问题. 因此，为了使药物开发过程更贴近于真实情况，有必要将定量的建模方法引入到药物研究领域，精确地模拟药物与靶标相互作用进而发挥药效的过程，发现更加有效的药物作用靶点.

随着实验技术的发展、数据的累积和文本挖掘的开展，生物通路的建模方法得到了快速的发展和应用. 其中，最常用的建模方法是确定性生化反应描述，已成功地用于药物代谢动力学和药剂反应建模. 确定性反应的缺点在于缺乏可伸缩性. 通常，基因组和蛋白质组方法要处理数十甚至数百个分子之间的信号网络，反应参数的范围可能包含多个跨度，超出了确定性方法的处理能力. 最新出现的方法，如结合反应(combinatorial reaction generation)和线性规划(linear programming)可以满足这种需求，批量地处理大规模的复杂化学反应网络. 进一步，随机方法能够从根本上克服确定性方法的限制. 它们是高度可伸缩性的，同时易于进行模拟. 然而，面对复杂的非线性动态问题，随机方法也存在很大的难度，还有待进一步探索.

近年来，用于描述反应动力学网络的数学模型被证明可以有效地预测生物体对于环境刺激和外界扰动的响应，识别可能的药物靶标[25]. 一种系统的药物设计方法是: 在网络中模拟单个反应的抑制过程，量化在指定观察量上的作用效果. 在代谢网络中，观察量一般是稳态值；在信号级联模型中，观察量包括浓度、特征时间、信号持续时间和信号幅值等. Schulz 等[26]在系统生物学建模语言(SBML)的基础上开发了一款名为 TIde 的工具，采用普通微分方程对系统进行模拟，研究在网络中不同位置进行激活和抑制处理时系统的响应. 通过模拟不同的抑制目标、类型和抑制剂浓度，确定一个或多个优化的药物靶标，在尽可能少的抑制剂数目下以较低的浓度使指定的观察量达到期望值. 此类药物作用模型的建立和模拟有助于理解药物的作用机制，预测药效发挥过程中可能存在的问题，进而为实验设计提供辅助作用.

**2.5.3** 多组学数据的综合应用.

系统生物学的优势在于“整合”，即综合利用基因组学、转录组学、蛋白质组学和代谢组学研究药物对系统的影响，提示可能的作用靶点. 例如，Chu 等[27]根据大规模实验及相关数据库建立了整合的蛋白质相互作用数据集，采用非线性随机模型、最大似然参数估计和 Akaike 信息准则(AIC, Akaike information criteria)方法，通过基因芯片数据估计疾病状态和正常状态下的蛋白质相互作用网络差异，识别受到扰动的枢纽(Hub)蛋白节点，发现候选的药物靶标. 除将转录组和蛋白质组数据结合之外，基因组与转录组、基因组与蛋白质组甚至更多组学数据的整合研究也在进行中.

整合研究的关键是以生物网络为中心加深对整个系统的理解. 疾病是一个非常复杂的生理和病理过程，涉及到多基因、多通路、多途径的分子相互作用的过程，这种网络化的特点对于药靶筛选至关重要. 系统生物学为药物开发过程提供了全新的视野，将蛋白质靶标置于其内在的生理环境中，在提供网络化的整体性视角的同时不会丧失关键的分子作用细节. 鉴于生物网络具有一定的冗余性和多样性，包括一定的反馈回路和故障安全(Fail-safe)机制. 因此，筛选潜在药靶时要考虑到其在网络中的位置，优先挑选那些处于枢纽位置发挥重要作用的靶点，并且避免反馈回路对药效进行补偿[28].

同时，疾病相关网络的内部高连接度表明，基于网络的诊疗方法应以整个通路作为靶标，而不是单个蛋白质. 其最终的目标不仅是识别一组能够共同发挥作用的药物，而且发现一组靶标或模块的组合，它们在不同的治疗位置发挥作用并最后集中到一个特定的通路位点. 尽管看起来这是一个几乎不可能实现的任务，但是在乳腺癌转移上的实验已经证明了基于通路知识进行多靶点联合治疗的有效性[29].

## 3 潜在药靶的生物信息学验证

在大量的潜在药靶被揭示之后，在此基础上可以寻找针对性的抑制小分子，进行后续的动物实验、临床测试等一系列药物开发过程．由于药物开发的难度较大、周期很长，在前期对候选药靶进行充分的筛选和验证显得非常必要．生物信息学方法在对候选药靶进行功能分析、预测其可药性并降低药物副作用方面也有重要的应用．

### 3.1 蛋白质的可药性

随着超过上百个真核和原核生物的基因组被完整测序，人们有机会对基因进行大规模的分析和筛选，据估计整个人类基因组中约有 10%与疾病相关，从而导致约 3 000 个潜在的药物靶标．同时，还有成千上万个来自于微生物和寄生生物的蛋白质，可以作为传染病治疗的药靶．目前，在所有的人类基因产物中仅有 2% (260～400)成功地发展为小分子药物的靶标．从大量的潜在靶标中挖掘能够被疾病修饰的可药部分是药物靶标验证的重要环节．

根据基因组信息和蛋白质结构特征，人们开发了一系列生物信息学方法预测潜在靶标的可药性．评估蛋白质可药性的第一步是识别在蛋白质表面的所有可能的结合位点，进而寻找真实的配体可结合位点[30]．其计算方法主要分为两类: 基于几何的方法和基于能量的方法．几何基础上的方法利用了这样一个事实: 天然的配体结合位点在蛋白质表面倾向于内部凹陷，例如 SURFNET、LIGSITE、SPROPOS、CAST、PASS 和 Flood-fill 方法．而能量基础上的方法将多种物理指标综合到 pocket 识别过程，试图计算其结合能，如 GRID、vdW-FFT、DrugSite 和 Computaional solvent mapping．在排序过程中，这些方法都能够给予真实的配体结合位点以较高的打分，证实了其有效性．第二步是评估结合位点能否高亲和性、特异地与小分子药物结合．定量评估给定位点可药性的计算工具较少，最直接的评估蛋白质可药性的方法是根据生物化学谱实际测量小分子击中目标的数目和类型，如 NMR 谱图．

此外，由于大部分的蛋白质是通过与其他蛋白质相互作用发挥生物学功能，蛋白质相互作用在组织的各种细胞过程中发挥了基础和关键作用，被认为是一种富于挑战的同时又充满吸引力的小分子药物作用的新型靶标．类似于单个蛋白质的可药性，人们提出了多种方法预测蛋白质相互作用的可药性[31−32]．2007 年，Sugaya 等[31]从 3 个方面评估蛋白质相互作用的可药性：蛋白质相互作用中包含的结构域对、蛋白质与小分子药物的结合位点、GO 功能注释的相似性打分．最近，Sugaya 等[32]使用结构、药物和化学以及功能相关的 69 个特征作为支持向量机的输入，判断 1 295 对已知结构的蛋白质相互作用的可药性，在标准的相互作用数据集中得到了 81%的预测准确率，其中区分度最大的特征是相互作用蛋白质的数目和通路数目．

### 3.2 药物的副作用

多组学数据的大量累积为药物研究提供了发展机遇，人们开发了多种方法用于发现潜在的药物靶标，但是最终找到合适的药物作用靶标并成功地进行临床应用并非易事．一般选择药物作用靶标要考虑两个方面的情况: 首先是靶标的有效性，即靶标与疾病确实相关，通过调节靶标的生理活性能够有效地改善疾病症状；其次是靶标的副作用，如果对靶标的生理活性的调节不可避免地产生严重的副作用，那么将其选作药物作用靶标也是不合适的．

药靶和药物代谢酶多态性是造成药物疗效差异和毒副作用的主要原因之一．药物反应个体差异与个体的基因多态性特别是单核苷酸多态性(single nucleotide polymorphism，SNP)密切相关．SNP 主要是指在基因组水平上由单个核苷酸的变异所引起的 DNA 序列多态性．SNP 在人类基因组中广泛存在，平均每 500～1000 个碱基对中就有 1 个，估计其总数可达 300 万个甚至更多．事先确定药物靶标的基因多态性，就可以估计药物适用的人群，进行个性化的医疗，增加疗效并降低毒副作用．目前，随着快速、规模化技术的发展，大量的 SNP 已经被揭示，为相关研究提供了基础．而生物信息学方法可以帮助阐释 SNP 与疾病治疗之间的关系，发现疾病易感基因和潜在药物靶标，评估药物疗效和毒副作用．以乳腺癌为模型，Wiechec 等[33]报道 SNP 基因型会影响 DNA 修复基因的转录活性和药物代谢过程，从而影响到临床的治疗毒性和效果．

进一步，在生物网络基础上综合评估药物作用的多种影响，也有助于寻找增加药物疗效、降低副作用的有效方法．在蛋白质 - 药物相互作用网络的基础上，Xie 等[34]介绍了一种新的计算策略识别基因组规模的蛋白质 - 受体结合谱，用于阐释 CETP 抑制剂的药物作用机制．通过将药物靶标与生物学通路相关联，揭示了 CETP 抑制剂的副作用受多个

交联通路的联合控制，给出了降低此类药物副作用的可能方法.

## 4 结论和展望

随着大规模组学数据的积累，仅凭实验方法已经不能满足数据分析和药靶发现的需求，有必要发展有效的生物信息学方法存储、分析、处理和整合多组学数据，提高药靶发现和验证的效率. 目前，生物信息学方法已成功地运用于药靶发现的各个环节，对于存储疾病相关的医学数据、发现大量潜在的药物靶标、揭示药物作用机理、评估作用靶点的可药性等方面做出了重要贡献，有利于设计更加有针对性的生物学实验，促进现代新药开发进程.

相比其他方法，采用生物信息学预测潜在药物靶标的优势在于：a. 不局限于特定的技术或某种类型的信息，尤其适合将不同的数据整合到一个大的体系中评估潜在药靶的表现；b. 以网络为基础的药靶发现平台有利于从整体角度进行药靶筛选并发现联合靶标；c. 随着动态的详细的生物学时空数据的累积，有可能在计算机中精确地模拟药物针对靶标作用的过程以及对整个系统产生的影响，从而大大提高药物开发的效率.

生物信息学方法在药物靶标发现的应用还刚刚起步，有赖于生物学理论、实验技术、统计分析和建模方法等多方面的进一步发展，从而在后基因组时代的疾病诊断、预后和个性化医疗中发挥更加重要的作用.

## 参 考 文 献


[1] Maglott D, Ostell J, Pruitt K D, *et al*. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res, 2007, **35**(Database issue): D26-D31

[2] Gao Z, Li H, Zhang H, *et al*. PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics, 2008, **9**: 104

[3] Rhodes D R, Kalyana-Sundaram S, Mahavisno V, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18 000 cancer gene expression profiles. Neoplasia, 2007, **9**(2): 166-180

[4] Le Novre N, Bornstein B, Broicher A, *et al*. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research, 2006, **34**(Database issue): D689-D691

[5] Olivier B, Snoep J. Web-based kinetic modelling using JWS Online. Bioinformatics, 2004, **20**(13): 2143-2144

[6] Ricke D O, Wang S, Cai R, *et al*. Genomic approaches to drug discovery. Curr Opin Chem Biol, 2006, **10**(4): 303-308

[7] Conde-Pueyo N, Munteanu A, Solé R V, *et al*. Human synthetic lethal inference as potential anti-cancer target gene detection. BMC Syst Biol, 2009, **3**: 116

[8] Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One, 2009, **4**(8): e6536

[9] Subramanian A, Kuehn H, Gould J, *et al*. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics, 2007, **23**(23): 3251-3253

[10] Autio R, Kilpinen S, Saarela M, *et al*. Comparison of Affymetrix data normalization methods using 6 926 experiments across five array generations. BMC Bioinformatics, 2009, **10**(Suppl 1): S24

[11] Stafford P, Brun M. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. Nucleic Acids Research, 2007, **35**(10): e72

[12] Rhodes D R, Yu J, Shanker K, *et al*. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA, 2004, **101**(25): 9309-9314

[13] Xu L, Geman D, Winslow R L. Large-scale integration of cancer microarray data identifies a robust common cancer signature. BMC Bioinformatics, 2007, **8**: 275

[14] Bakheet T M, Doig A J. Properties and identification of human protein drug targets. Bioinformatics, 2009, **25**(4): 451-457

[15] Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 2006, **22**(22): 2800-2805

[16] Ostlund G, Lindskog M, Sonnhammer E L. Network-based identification of novel cancer genes. Mol Cell Proteomics, 2010, **9**(4): 648-655

[17] Li L, Zhang K, Lee J, *et al*. Discovering cancer genes by integrating network and functional properties. BMC Medical Genomics, 2009, **2**: 61

[18] Hu X, Zhang Y, Zhang A, *et al*. Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls *via* label-free semiquantitative shotgun technology. OMICS, 2009, **13**(4): 291-300

[19] Sleno L, Emili A. Proteomic methods for drug target discovery. Curr Opin Chem Biol, 2008, **12**(1): 46-54

[20] Pohjanen E, Thysell E, Lindberg J, *et al*. Statistical multivariate metabolite profiling for aiding biomarker pattern detection and mechanistic interpretations in GC/MS based metabolomics. Metabolomics, 2006, **2**(4): 257-268

[21] Sridhar P, Song B, Kahveci T, *et al*. Mining metabolic networks for optimal drug targets. Pac Symp Biocomput, 2008, **13**: 281-302

[22] Li Z, Wang R S, Zhang X S, *et al*. Detecting drug targets with minimum side effects in metabolic networks. IET Syst Biol, 2009, **3**(6): 523-533

[23] Yang Y, Adelstein S J, Kassis A I. Target discovery from data mining approaches. Drug Discov Today, 2009, **14**(3-4): 147-154

[24] Wang H, Huang M, Zhu X. Extract interaction detection methods from the biological literature. BMC Bioinformatics, 2009, **10**(1): S55

[25] Purohit R, Rajendran V, Sethumadhavan R. Relationship between mutation of serine residue at 315th position in *M. tuberculosis* catalase-peroxidase enzyme and Isoniazid susceptibility: An in silico analysis [J/OL]. J Mol Model [2010-07-01]. http://www.springerlink.com/content/cb31324278mu8841/

[26] Schulz M, Bakker B M, Klipp E. Tide: a software for the systematic scanning of drug targets in kinetic network models. BMC Bioinformatics, 2009, **10**: 344

[27] Chu L H, Chen B S. Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. BMC Syst Biol, 2008, **2**: 56

[28] Zanzoni A, Soler-López M, Aloy P. A network medicine approach to human disease. FEBS Lett, 2009, **583**(11): 1759−1765

[29] Pujol A, Mosca R, Farrés J, *et al*. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci, 2010, **31**(3): 115−123

[30] Halgren T A. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model, 2009, **49**(2): 377−389

[31] Sugaya N, Ikeda K, Tashiro T, *et al*. An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacol, 2007, **7**: 10

[32] Sugaya N, Ikeda K. Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinformatics, 2009, **10**: 263

[33] Wiechec E, Hansen L L. The effect of genetic variability on drug response in conventional breast cancer treatment. Eur J Pharmacol, 2009, **625**(1−3): 122−130

[34] Xie L, Li J, Xie L, *et al*. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLoS Comput Biol, 2009, **5**(5): e1000387

# Potential Drug Target Discovery Based on Bioinformatics Methods*

LIU Wei, XIE Hong-Wei**

(*Department of Automatic Control, College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, Changsha* 410073, China)

**Abstract** Typically a drug target is a key molecule involved in a particular metabolic or signaling pathway, that is specific to a disease condition or pathology. Drugs may be designed that bind to the active region and inhibit this key molecule. Determining specific disease-related target molecules is the basis of modern drug development. In the process of drug target discovery, bioinformatics methods play irreplaceable roles, especially suited for the analyses of large-scale and multi-omics data. On current, many disease-related database resources have emerged. Various bioinformatics methods have been established based on biological network characteristics, multiple gene chips, proteomics and metabolomics data to discover potential drug targets, and predict the target druggability and side effects of drugs.

**Key words** drug target, network feature, gene microarray, proteomics

**DOI**: 10.3724/SP.J.1206.2010.00251

*This work was supported by grants from National Basic Research Program of China (2006CB910803, 2006CB910706), Hi-Tech Research and Development Program of China (2006AA02A312), National S & T Major Project (2008ZX10002-016, 2009ZX09301-002), The National Natural Science Foundation of China (60773021, 60603054, 30621063, 30800200) and State Key Laboratory of Proteomics (SKLP-Y200811).

**Corresponding author.

Tel: 86-731-84573369, E-mail: xhwei65@nudt.edu.cn

Received: May 10, 2010 Accepted: August 18, 2010