文章编号：1004-0374(2009)03-0400-08

应用化学基因组信息预测小分子化合物的潜在生物靶标的理论方法

李 嫣，王任小*
（中国科学院上海有机化学研究所 生命有机化学国家重点实验室，上海 200032）

摘 要：在后基因组时代，化学基因组技术在药物作用靶点的确认、小分子化合物对通路的作用，以及小分子先导化合物的识别等方面都有着广泛的应用，为新药研发提供了新的技术方法。本文主要介绍了当前几种基于化学基因组信息来预测小分子化合物潜在生物靶标的理论方法(包括化学相似性搜索方法、反向分子对接方法、数据挖掘方法以及生物活性谱图分析方法)，并分析了这些方法的优缺点以及应用前景。

关键词：化学基因组；生物靶标预测；数据挖掘；生物活性谱图
中图分类号：Q78； Q812 **文献标识码**：A

Theoretical approaches to the prediction of the biological targets of small-molecular compounds based on chemogenomic information

LI Yan, WANG Ren-xiao*
(State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China)

Abstract: In this post-genomic era, chemogenomics can be applied in target elucidation, understanding of the effects of small-molecular compounds on biological pathways, and discovery of novel active compounds. These new techniques collectively play an important role in modern drug discovery. This article reviews the existing theoretical approaches to the prediction of biological targets of small-molecular compounds based on chemogenomic information, including chemical similarity searching, reverse docking, data mining, and bioactivity spectrum, and depicts the strength and shortcomings of these methods as well as their perspectives in the future.

Key words: chemogenomics; target elucidation; data mining; bioactivity spectrum

1 引言

人类基因组计划(human genome project, HGP)的完成揭示了人类基因组所包含的约20 000 － 25 000个基因[1]。根据这些基因数目推测可用于治疗人类疾病的潜在药物靶标大约有2 000 － 5 000种[2]。而在过去几个世纪中人们发现并用于药物研发的靶标总数仅约500个[3]。因此，数量庞大的潜在靶标尚未得到功能确证以及三维结构测定。如何在缺少生物靶标确切信息的情况下快速有效地确认出这些潜在靶标呢？伴随着这个问题的提出，一个新兴的研究领域——化学基因组学(chemogenomics)应运而生。

化学基因组学最初被定义为一种基于基因家族的药物发现方法，它“用来描述对所关注的靶标基因家族的探索，即利用这一家族中某一已知成员的小分子先导化合物来研究其他未知成员的生物功能”[4]。在实际应用中，化学基因组学的关注对象

收稿日期：2008-12-09；**修回日期**：2009-01-12
基金项目：国家自然科学基金(20502031, 20772149, 90813006)；“863”项目(2006AA02Z337)； 上海市科委项目(074319113)
***通讯作者**：Tel：021-54925128； E-mail：wangrx@mail.sioc.ac.cn

已经不再局限于基因，它可以运用各种技术手段来研究小分子化合物在基因、蛋白，甚至组织器官水平上的生物响应。这些生物响应可以通过基因表型输出或高通量筛选技术测量获得。这些表观的生物响应信息不仅可以用于阐明疾病的生理机制，还可以从中推测出小分子化合物潜在的生物作用靶标。针对小分子化合物在多种生物实体上体现出来的生物效应的综合研究可以为药物设计提供更多的信息，从而提高药物设计的成功率。化学基因组学作为后基因组时代的新技术，它可以弥补目前基于单靶标的药物设计过程中忽略其他潜在靶标的影响这一缺陷。这种新的研发模式有望大大促进新药研发过程。

化学基因组学的重要应用就是根据已有的各种生物和化学信息来预测有机化合物的未知作用靶标，预测给定化合物的生物活性等，提高新药研发的效率。本文结合国内外多个课题组的研究成果，向读者主要介绍利用化学基因组信息来预测有机小分子化合物的潜在作用靶标的理论方法。

2　应用化学基因组信息预测有机小分子化合物作用靶标的计算方法

目前用于预测有机小分子化合物作用靶标的理论方法大致上可以分为四大类[5,6]：化学相似性搜索方法、反向分子对接方法、数据挖掘方法和生物活性谱图(bioactivity spectrum)分析方法。对前两种方法我们将进行简单的介绍，后两种方法属于典型的化学基因组学方法，我们将结合国内外的研究成果对其进行比较详细的介绍。

2.1　化学相似性搜索方法　化学相似性搜索是一种广泛应用于生物靶标预测的计算方法。它所依据的原理就是结构或化学性质相似的小分子化合物对应于性质相同或相近的靶标[7]。因此，可以通过比较给定分子与化合物数据库中已知作用靶标的小分子的结构或化学性质来预测给定分子的潜在作用靶标。用于进行相似性比较的描述符可以是一维、二维或三维的，其中二维描述符因其较高的计算效率而经常被应用于靶标预测。常用的二维描述符采用基于指纹方法生成的拓扑描述符，如MDL Public Keys、SciTegic ECFP(extended connectivity fingerprints)等。相似度的计算方法也有很多。最常见的是Tanimoto系数，其计算公式为：S_T=C/(A+B-C)，A和B分别为化合物A和B中所定义特征结构的数目，C为两个化合物中共有的特征结构的数目。图1中的范例展示了多种分子描述符的生成方法[8]。

Nettles等[9]分别采用二维描述符MDL Public Keys、ECFP和三维描述符FEPOPS(FEature POint PharmacophoreS)进行基于化学相似性搜索的靶标预测。他们收集了WOMBAT 2005数据库中47 505个结构特异的活性化合物和它们对应的544个靶标，

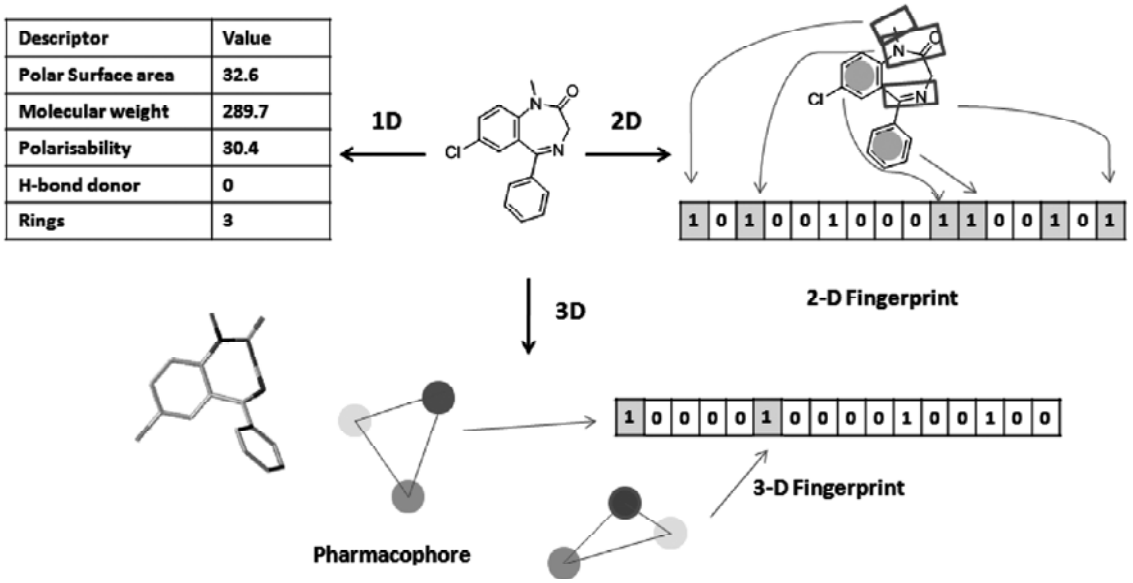

图1　有机小分子化合物的描述符示例[8]

一维描述符：依次为小分子的极性表面积、分子量、极化率、氢键给体数目和环的数目；二维描述符：根据分子特征结构编码的指纹，这里选取的特征结构包括苯环、C=N基团、酰胺基团和-NCH_3基团；三维描述符：根据小分子三维结构产生的药效团模型进行指纹编码

这些靶标每个至少对应有两个活性化合物，每个化合物可以看成是一个已知靶标的探针分子。他们分别采用二维和三维描述符进行化合物之间的两两相似性比较，计算它们的Tanimoto系数。与探针分子相似度最高的化合物作为参照分子，用于靶标的预测。结果显示采用二维描述符进行相似性计算的预测成功率明显高于三维描述符。而对于与探针分子的二维描述符相似度较低的化合物，三维描述符则更适合用于预测其靶标。

基于结构相似性搜索的方法非常迅速，数秒之内可以获得大量的反馈结果，它不要求数据库必须具备标准化的靶标命名，因此任何一个含有靶标注释信息的化合物数据库都可以利用相似性搜索方法来进行给定分子的靶标预测。近些年来，随着计算机技术的发展，可以通过网络访问的化合物数据库大多提供了结构相似性搜索功能，使得该类方法的应用更加普遍。但是该方法也存在一些问题：如何从获得的大量预测结果中进行选择；出现频率高的靶标如何考虑其优势。结构相似性搜索方法仅考虑了小分子化合物的化学性质和结构信息，它们与靶标之间的相互作用信息并没有在相似性搜索中充分体现出来。

2.2 反向分子对接方法 分子对接方法[10]通常用于研究若干小分子化合物与给定生物靶标分子(蛋白或核酸等生物大分子)的结合。顾名思义，反向分子对接方法则是将某给定小分子化合物与若干个生物靶标分子进行分子对接，从中挑选出结合情况最好的候选者，认为其有可能就是给定小分子化合物的生物靶标分子。

Chen和Zhi[11]第一个成功地将反向分子对接方法应用于药物分子4H-三苯氧胺和维生素E的靶蛋白预测(这两种药物分子都具有多个生物靶标)。他们选择来自人和哺乳动物的蛋白质分子作为候选靶标，所有蛋白分子的三维结构均从PDB数据库获得。根据分子对接程序DOCK中的算法[12]，他们通过一组可重叠的球体定义蛋白分子可能的结合位点，总共获得了2 700个定义结合位点的蛋白分子结构。通过药物分子与这些蛋白结构的反向分子对接来预测它们的潜在靶蛋白。反向分子对接过程中采用INVDOCK程序，对接得到的复合物结构采用基于蛋白-配体相互作用能的亲合性打分函数进行评价，主要考虑蛋白-配体之间的氢键作用和非共价作用两项。对预测靶标的评估不仅考虑了已知药物分子与蛋白分子的亲合能，还与其他配体分子与同一蛋白的亲合能进行比较来分析药物分子的竞争性结合能力。最终预测获得了这两个药物分子的一系列可能的作用靶标，其中大约有50%的蛋白为已确认的药物作用靶标或通过实验获得了验证。Li等[13]发展了类似的基于反向分子对接方法的“靶标垂钓”工具——TarFisDock。他们根据该方法预测肽脱甲酰基酶(peptide deformylase)有可能是来自中草药紫金标中具有抗幽门螺旋杆菌性质的有效成分的潜在靶标。这一预测随后通过酶活性抑制测定、X-晶体衍射结构等实验方法得到了证实[14]。

虽然通过反向分子对接方法预测靶标不乏成功之例，但是该方法的推广不容乐观。首先，该方法只能考虑已知三维结构或通过同源模建等方法可以可靠预测结构的生物靶标分子。具有明确结构信息的靶标分子目前只占所有潜在靶标分子的一部分。其次，通过反向分子对接推测生物靶标极大地依赖于分子对接方法的精度。分子对接方法本身以及所依赖的打分函数的精度仍需要提高[15,16]，并且需要考虑亲合性得分在不同蛋白体系上如何归一化的问题。从技术层面上来看，批量进行反向分子对接耗用计算资源较多，所需的准备工作以及对计算结果的分析都较为繁琐。以上因素都影响了该方法在现阶段的推广。

2.3 基于注释化学数据库的数据挖掘方法 基因芯片等高通量技术的发展和应用可以大批量地产生多种类型的生物活性数据。要从这些海量的数据中提取有用的信息，则必须依赖于有效的数据挖掘手段。机器学习是常用的数据挖掘方法之一[17]。它要求使用一部分数据作为训练集，然后通过自动学习来构建合适的预测模型。目前国内外很多的科研单位和公司都构建了包含标准化注释信息的小分子数据库(也称为化学基因组数据库)。这些数据库为建立预测小分子化合物作用靶标的数据挖掘方法提供了很好的素材。表1和表2列出了部分常用的此类数据库[5,18,19]。这些数据库中的信息大多收集自公开发表的化学或生物学期刊和专利，不仅提供了小分子化合物的化学结构信息，也收集小分子对应的靶蛋白及相应的活性数据(如K_d、IC_{50}等)。根据这些分子结构信息和生物活性数据所建立的模型不仅可以用来预测小分子化合物的主要靶标，还可以预测它的次级靶标以及在临床上的副作用等，从而更全面地评价小分子化合物成药的可能性。

Niwa[20]从MDL MDDR数据库中选出799个在七大类生物靶标上显示出活性的小分子化合物，构

表1　可用于数据挖掘方法预测生物靶标的商业数据库

Databases	Companies and their web links	Contents
Target inhibitor database	GVK Bio: http://gvkbio.com/informatics.html	1.8 M entries, 500K compound records, 1.5K targets
MedChem (GVK Bio)	GVK Bio: http://gvkbio.com/informatics.html	750K compound records, 607K unique compounds, 4900 targets
AurSCOPE (Aureus)	Aureus: http://www.aureus-pharma.com/	GPCR: 152K compounds, 635K activities; Kinases: 51.8K compounds, 163.7K activities; Ion channel: 58.4K compounds, 217.6K activities
stARLITe	Inpharmatica: http://www.inpharmatica.co.uk/	300K compounds, ~5000 targets, 1.3M data
ChemBioBase Suite	JubilantBioSys: http://www.jubilantbiosys.com/	~1020 targets. Kinases: 319K compounds; GPCR: 400K compounds; Nuclear receptor: 150K compounds; Ion Channel: 100K compounds; Protease: 400K compounds
BioPrint	Cerep: http://www.cerep.fr/	180 diverse targets, 2500 drugs, >1 M records, in vivo data, adverse effects
WOMBAT	SunsetMolecular: http://sunsetmolecular.com/	154K entries, 136K unique compounds, 308K total activity data, 1320 protein targets
MDDR	MDL: http://www.mdl.com/	~160K entries, 123.7K unique compounds, ~700 targets, bioactivity data, chemical classes

表2　一些包含药物-靶标作用信息的公开数据库

Database	Web links	Content
DrugBank	http://www.drugbank.ca/	~1000 FDA-approved drugs; ~3000 experimental drugs; 6000 drug-targets relationships; chemical, pharmacological and pharmaceutical data
Matador	http://matador.embl.de/	~770 drugs; ~7000 direct and ~5000 indirect drug-target relationships; links to literature sources for interactions
SuperTarget	http://insilico.charite.de/supertarget/	~1500 drugs; 7300 drug-target relations
Therapeutic Target Database (TTD)	http://bidd.nus.edu.sg/group/cjttd/TTD_ns.asp	~2100 drugs; Drug-target relationships with 1535 targets
PDSP K_i	http://pdsp.med.unc.edu/pdsp.php	~6800 chemicals; ~46000 K_i values
Binding DB	http://www.bindingdb.org/	~18000 chemicals; ~30000 records with K_i, IC_{50}, or thermodynamic data
PubChem BioAssay	http://pubchem.ncbi.nlm.nih.gov/	~560000 chemicals; ~600 single compound and high-throughput screening assays
ChemBank	http://chembank.broad.harvard.edu/	~1.2 million chemicals; 2500 high-throughput biological assays from 188 screening projects
NCI tumor cell line database	http://dtp.nci.nih.gov/webdata.html	~43000 compounds with screening data on 60 tumor cell lines, mRNA expression data
PDBbind-CN	http://www.pdbbind.org.cn	~3600 protein-ligand complexes with known binding data; ~720 protein-protein and protein-nucleic acid complexes with known binding data; ~8700 small-molecule ligands

建成一个数据集，并随机选取其中60%用作训练集，20%为用于改进模型参数的测试集，剩余20%为用于评估模型预测能力的预测集。在预测模型的建立过程中，他仅以包含C、H、N、O、S、P和卤素等元素在内的24种原子类型作为化合物结构的描述符，并结合概率神经网络方法对具有靶蛋白注释信息的化合物进行学习。图2中给出了一个简单二类分割问题的概率神经网络结构。它共分为四

层：输入层对应于化学描述符X；隐含层代表训练模式，它所包含的结点数目等于训练集中化合物的总数；合计层所包含的结点数目则等于所划分的靶标类别总数；输出层给出化合物对应于某一靶标类别的概率 f。将训练集中的化合物描述符和已知的靶标类别作为初始输入和输出在该网络模型上进行多次学习来调整获得最优参数，然后利用该模型预测未参与训练的化合物的靶标。最终的预测结果显示67% — 98%的化合物被正确地划分到了所属的靶蛋白家族中。

Nidhi等[21]则结合多类别的Naïve Bayesian模型对来自WOMBAT 2005数据库的964个已知靶蛋白的活性化合物进行训练，建立预测模型。在多类别的Bayesian模型中，每个靶蛋白对应于一个类别，使用二维化学描述符ECFP作为区分活性化合物和非活性化合物的特征。最终模型输出每类别靶蛋白的

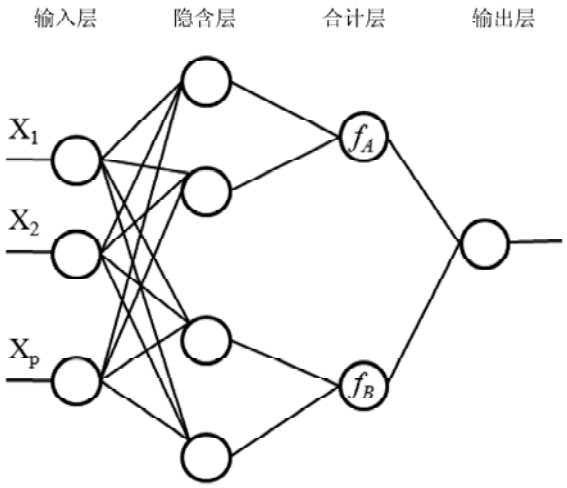

图2 概率神经网络结构示意图

Bayesian得分，得分越高则说明该类蛋白成为输入化合物的对应靶标的可能性越大。该模型的原理与概率神经网络类似，只是所采用的数学方法不同。用该模型对MDL MDDR数据库中分属10个类别靶标的活性化合物进行预测，77%的化合物预测得到了正确类别的生物靶标。对于仅提供疗效或基因水平活性信息的化合物也同样预测到了与之对应的靶蛋白类别。

Hert等[22]以取自注释化学数据库MDDR和WOMBAT中的靶标-配体数据集为基础，建立了基于配体分子化学结构相似性的化学信息学网络。网络的节点为每个靶标所对应的配体分子，这种网络图可用于预测未知配体分子的可能作用靶标。与基于蛋白序列相似性的生物信息学网络相比，化学信息学网络以配体为中心衡量靶标之间的相似性，它能够揭示一些仅通过生物信息(如序列相似性)无法预测的内在联系。他们的研究还表明在采用不同的分子描述符以及不同的相似性计算方法时能够获得稳定的化学信息学网络。这些优势引起了人们对该方法的兴趣。Yamanishi等[23]在研究中同时考虑了药物分子化学结构和蛋白序列的相似性，提出通过整合化学和基因学空间的策略来预测靶标-药物的作用网络。他们采用双向图学习方法(bipartite graph learning)建立药物-靶标作用网络模型，该模型因药物分子与靶标两方面信息的整合而提高了预测的准确率，相对于Hert等[22]的研究有了进一步的拓展。

基于数据挖掘的理论方法快速灵活，精确度比较高，可以应用于研究多样性的化合物，而且可以广泛地整合于各种含注释信息的化学数据库。可采用的机器学习方法有很多种，目前比较流行的有Bayesian模型和支持向量机模型等[24,25]。机器学习方法的主要缺点是，首先，需要一个已知训练集来建立预测模型，因此无法对训练集之外结构差异较大的目标化合物进行预测。其次，所采用的训练集必须要求具有精确的注释信息，即小分子与靶标有明确的对应关系而且靶标的命名需要标准化，因此普通的化合物数据库并不适用。再次，由于可以使用的生物活性数据的来源以及类型都不统一，此类方法一般不能进行定量的预测。

2.4 生物活性谱图分析方法 化合物在某一系列细胞模型或者蛋白分子上所表现出来的生物活性数据的总合就构成了该化合物的生物活性谱图。这种生物活性谱图反映了小分子化合物对多个生物靶标的生物效应，更全面地体现了小分子的药理性质。因此对这类数据的分析可以为药物设计提供重要的信息，提高新药研发的成功率。

NCI60抗肿瘤药物筛选数据库给出了超过43 000种化合物在60种肿瘤细胞系上的GI_{50}值(细胞增殖半数抑制浓度)。这一系列GI_{50}值形成了每种化合物在这些肿瘤细胞系上的生物活性谱图。依据NCI60抗肿瘤药物筛选数据库进行研究并得以公开报道的有很多，包括依据化合物活性谱图对肿瘤细胞系筛选数据的聚类分析，并以此为模型根据化合物的化学结构相似性预测给定化合物的可能靶标以及在各个细胞系上的活性[26,27]；研究化合物高通量筛选结果与mRNA表达水平的相关性[28]或者化合物抑制肿瘤细胞的作用机制与化合物活性谱图之间的

相关性[29]；化合物对不同细胞系的化学敏感性以及在未知细胞系上的生物活性预测等[30]。

Cerep公司出品的BioPrint数据库则提供了在单一浓度(10 μmol/L) 下1 567种类药化合物作用于不同靶标蛋白的活性抑制百分比。相对于NCI60 在细胞系上获得的生物活性数据，该数据库提供了小分子直接与不同蛋白的相互作用信息，可以明确地将小分子的药理性质与具体的靶标蛋白挂钩。辉瑞全球研发中心的Fliri等应用该数据库进行了大量的研究工作[31,32]。他们对1 567个化合物的生物活性谱进行分层聚类分析，采用聚类相似置信值(CCS)来定量计算谱图之间的相似度。CCS 值根据 cosine 相似系数计算获得，计算公式为 $S_c=C/(A*B)^{1/2}$，式中 A 和 B 分别为谱图 A 和 B 中所定义特征指纹的数目，C 为两张谱图中共有的特征指纹的数目。假设横坐标为 92 种蛋白分子，纵坐标为 1 567 个化合物，那么沿纵坐标聚类的结果显示了生物活性谱图与化合物结构的相关性，化学结构相似的化合物对应于相同或相似的靶蛋白；沿横坐标聚类的结果则揭示了蛋白结合位点的差异。图 3 为对一张生物活性谱双向分层聚类分析的示意图。通过生物活性谱图的相似性比较就可以预测化合物的分子药理性质。他们比较了化合物Ticonazole和Clotrimazole的生物活性谱图，计算得到CCS 值为 0.79。事实上两者确实都具有抗菌活性。通过化学结构的相似性比较也可以预测未知化合物的活性谱图，进而推测其可能发生作用的靶蛋白。

这种活性谱图的概念拓展至基因水平就是小分子化合物的基因表达谱图。小分子化合物的基因表达谱图就是表征该化合物对某一细胞系中各种基因

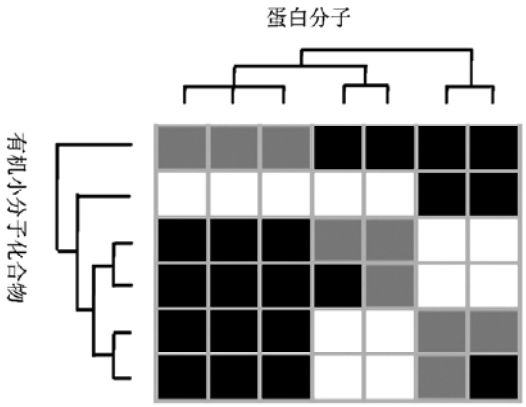

图3　生物活性谱图的双向分层聚类示例

纵轴代表有机小分子化合物；横轴代表蛋白分子；每个格点表示单一浓度下某一化合物对某个蛋白分子的抑制作用，以百分比表示，颜色越深抑制作用越强

的表达水平的影响，通常根据 mRNA 芯片所提供的信息来判断。研究小分子化合物的化学结构与其基因表达谱图之间的相关性已经成为后基因组时代药物研究的一项重要内容。麻省理工的 Lamb 小组收集了若干小分子化合物影响人类细胞系中基因表达水平的 mRNA 芯片数据，并对其进行了系统的研究[33]。他们发展了一套全局表达模式的比较方法，并建立了相应的搜索系统，称为“联络图”(connectivity map)——图 4 左侧显示的是所关注的小分子化合物在人类细胞系上的基因表达谱；图中间是数据库中不同化合物在人类细胞系上的参照谱图，他们发展了一种基于 Kolmogorov-Smimov 统计原理的非参数排名方法来对未知谱和参照谱进行相似性比较，它们之间的相似度由Connectivity Score来衡量; 图右侧显示的是参照谱图按最终得分从高到低的排名结果以及相对应的小分子化合物。从该结果可以获得与所关注化合物的基因谱图相似的已知化合物，进而预测其可能的作用靶标以及所影响的通路、相关的疾病等。他们根据该方法发现抗癌药物 Gedunin 可以作为 HSP90 蛋白的抑制剂。最近，Li 等[34]在 Lamb 的方法上进一步改进，发展了一种以基因表达模块为单元的功能相似性搜索方法。他们采用了基因家族的分类信息，每一类基因家族作为一个单元，对每个单元内的基因表达谱相关信息富集获得针对每个基因表达模块的生物活性谱图。这样可以将化合物对基因表达水平的影响与基因功能模块直接联系起来，获得更直观的结果。同时采用这种信息富集的方法也降低了实验数据的噪音和边际效应对预测结果的影响。

药物分子的副作用是一个非常复杂的表观现象，它通常与药物分子的“脱靶”效应(off-target effect)有关，反映了药物与其他次级靶标之间的相互作用。这也可以视为一种拓展的生物活性谱图。Campillos等[35]就提出了依据药物分子副作用的相似性来预测其对应靶标的策略。他们通过文本挖掘技术从药物说明书中提取药物分子的副作用信息，并以规范化的术语来表示。药物分子副作用的相似性计算结合了权重算法和统计显著性评估方法。对包含502个药物分子和4 857个已知靶标的训练集的研究表明，药物分子副作用的相似性与靶蛋白的共享具有一定的相关性。以训练集中的药物分子和靶标为参照，对 746 个上市药物进行两两比较后新发现有 261 对化学结构不相似的药物分子之间有着类似的药物副作用表现，也即它们有可能对应相同的靶

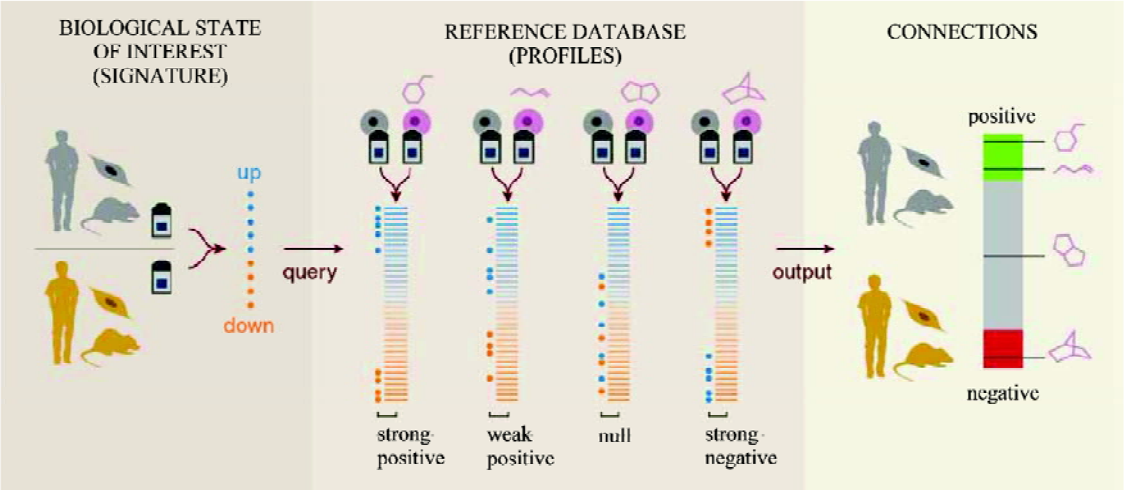

图4 “联络图”的使用流程示意图[33]

图左侧为所研究的化合物在人类细胞系上的基因表达谱；中间为数据库中收录的各种化合物的基因表达谱作为参照；图右侧显示各参照谱与未知谱进行相似性比较后按Connectivity Score排序的结果，从而推测所研究的化合物的生物活性

标。他们对其中20对药物分子对应关系进行了实验验证，最终通过体外结合测试及细胞水平的活性测试证实了13对药物-靶标关系。这也体现了生物活性谱图分析方法在结构多样的小分子化合物的靶标预测中的优越性。

生物活性谱图分析方法也存在一些固有的缺陷：在建立预测模型时要求采用在一系列生物实验中获得的完整数据。这样的数据来源比较匮乏，而且往往局限于某个特定的研究领域，如蛋白激酶的活性测试、细胞模型上的毒性测试等。因此，此类方法目前的应用范围比较有限。

3 总结与展望

本文主要介绍了四大类预测小分子化合物的潜在生物靶标的理论方法。化学相似性搜索方法基于小分子化学结构的描述符并结合已知的生物信息来预测化合物对应的靶蛋白，但是在确定靶蛋白的优先次序方面尚无系统性的方法。反向分子对接方法在结构层面上研究小分子化合物与多个蛋白之间的相互作用，但是该方法的应用受限于蛋白分子的三维结构信息以及分子对接方法的精度。应用化学基因组信息的预测方法则根据含有注释信息的化学数据库或通过高通量筛选所获得的生物活性谱图，采用机器学习或统计方法来归纳这些信息，弥补了前两种方法的不足，可以有效地进行靶标预测。基于注释化学数据库的方法采用化学描述符(如ECFP)作为分辨化合物在不同体系上所产生的生物响应的特定表征，而生物谱图分析方法则采用了复杂的生物描述符，也称生物指纹(biological fingerprint)。相对于化学结构，后者可以更为精确地反映出化合物与其生物效应的对应关系，逐渐成为了该领域中的主要发展趋势。但是目前生物谱图分析方法受限于生物数据来源，其应用局限在特定的领域中(如基因表达、肿瘤细胞系等)。因此，我们认为在研究实践中应该注意结合运用以上各类理论预测方法，取长补短。基于化学基因组信息的靶标预测方法不仅能够预测与小分子化合物作用的潜在生物靶标，还可以预测小分子化合物对生命体系中各组分的影响，结合已知的生物信息可用于对通路作用机制的探索。所有这些为药物分子设计提供了更丰富的构效关系知识，对于创新药物的研制和开发具有重要的意义。

[参 考 文 献]

[1] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 2004, 431(7011): 931-45

[2] Drews J, Ryser S. Human disease – from genetic causes to biological effects [M]. Berlin: Blackwell, 1997: 5-9

[3] Drews J, Ryser S. The role of innovation in drug development. Nat Biotechnol, 1997, 15(13): 1318-9

[4] Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet, 2004, 5(4): 262-75

[5] Jenkins JL, Bender A, Davies JW. *In silico* target fishing: predicting biological targets from chemical structure. Drug Discov Today Technol, 2006, 3(4): 413-21

[6] Bender A, Young DW, Jenkins JL, et al. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. Comb Chem High Throughput Screen, 2007, 10(8): 719-31

[7] Schuffenhauer A, Floersheim P, Acklin P, et al. Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci, 2003, 43(2): 391-405
[8] Rognan D. Chemogenomic approaches to rational drug design. Br J Pharmacol, 2007, 152(1): 38-52
[9] Nettles JH, Jenkins JL, Bender A, et al. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. J Med Chem, 2006, 49(23): 6802-10
[10] Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput Aided Mol Des, 2002, 16(3): 151-66
[11] Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins, 2001, 43(2): 217-26
[12] Kuntz ID, Blaney JM, Oatley SJ, et al. A geometric approach to macromolecule-ligand interactions. J Mol Biol, 1982, 161(2): 269-88
[13] Li HL, Gao ZT, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res, 2006, 34: W219-24
[14] Cai JH, Han C, Hu TC, et al. Peptide deformylase is a potential target for anti-*Helicobacter pylori* drugs: reverse docking, enzymatic assay, and X-ray crystallography validation. Protein Sci, 2006, 15(9): 2071-81
[15] Wang RX, Lu YP, Fang XL, et al. An extensive test of 14 scoring functions using the PDB bind refined set of 800 protein-ligand complexes. J Chem Inf Comput Sci, 2004, 44 (6): 2114-25
[16] Warren GL, Andrews CW, Capelli AM, et al. A critical assessment of docking programs and scoring functions. J Med Chem, 2006, 49(20): 5912-31
[17] Adams N, Schubert US. From data to knowledge: chemical data management, data mining and modeling in polymer science. J Comb Chem, 2004, 6(1): 12-23
[18] Kuhn M, Campillos M, Gonzólez P, et al. Large-scale prediction of drug-target relationships. FEBS Lett, 2008, 582 (8): 1283-90
[19] Wang RX, Fang XL, Lu YP, et al. The PDBbind database: methodologies and updates. J Med Chem, 2005, 48(12): 4111-9
[20] Niwa T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. J Med Chem, 2004, 47(10): 2645-50
[21] Nidhi'Glick CM, Davies JW, et al. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model, 2006, 46(3): 1124-33
[22] Hert J, Keiser MJ, Irwin JJ, et al. Quantifying the relationships among drug classes. J Chem Inf Model, 2008, 48(4), 755-65
[23] Yamanishi Y, Araki M, Gutteridge A, et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics, 2008, 24(13), i232-40
[24] Teschendorff AE, Wang YZ, Barbosa-Morais NL, et al. A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. Bioinformatics, 2005, 21(13): 3025-33
[25] Winters-Hilt S, Yelundur A, McChesney C, et al. Support vector machine implementations for classification & clustering. BMC Bioinformatics, 2006, 7(Suppl 2): S4
[26] Rabow AA, Shoemaker RH, Sausville EA, et al. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. J Med Chem, 2002, 45(4): 818-40
[27] Wang H, Klinginsmith J, Dong X, et al. Chemical data mining of the NCI human tumor cell line database. J Chem Inf Model, 2007, 47(6): 2063-76
[28] Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet, 2000, 24(3): 236-44
[29] Covell DG, Wallqvist A, Huang R, et al. Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small molecular screening and structural databases. Proteins, 2005, 59(3): 403-33
[30] Lee JK, Havaleshko DM, Cho H, et al. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proc Natl Acad Sci USA, 2007, 104(32): 13086-91
[31] Fliri AF, Loging WT, Thadeio PF, et al. Biological spectra analysis: linking biological activity profiles to molecular structure. Proc Natl Acad Sci USA, 2005, 102(2): 261-6
[32] Fliri AF, Loging WT, Thadeio PF, et al. Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. J Med Chem, 2005, 48(22): 6918-25
[33] Lamb J, Crawford ED, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006, 313(5795): 1929-35
[34] Li Y, Hao P, Zheng SY, et al. Gene expression module-based chemical function similarity search. Nucleic Acids Res, 2008, 36(20): e137
[35] Campillos M, Kuhn M, Gavin AC, et al. Drug target identification using side-effect similarity. Science, 2008, 321(5886): 263-6