PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(10): 923~935
www.pibb.ac.cn

面向药物发现和精准医疗的基因表达谱分析 *

刘 阳[1] 白 卉[2]** 陶 欢[3] 何 松[4] 黄 昕[4] 伯晓晨[4]** 王升启[4]**

([1] 中国人民解放军第三 O 二医院，北京 100039；[2] 中国人民解放军第 451 医院，西安 710054；
[3] 中国人民解放军第四军医大学，西安 710032；[4] 军事医学科学院放射与辐射医学研究所，北京 100850)

摘要 作为功能基因组学中重要的组成部分，基因表达谱在生物学、医学和药物研发等多个领域发挥着重要作用．特别是随着精准医疗概念的提出，整合多组学数据用于个性化医疗是未来的发展趋势．本文从基因表达谱的基本概念出发，重点介绍面向药物发现的基因表达谱分析方法，即基于关联图谱的方法、基于基因调控网络的方法和基于多组学数据整合的方法．系统整理了各种方法的研究进展，特别是在抗癌药物研发领域的最新进展，为利用基因表达谱数据进行药物研发提供方法借鉴.

关键词 基因表达谱，关联图谱，整合网络细胞印记库，基因调控网络，多组学融合，精准医疗
学科分类号 Q811.4，R9 **DOI**: 10.16476/j.pibb.2016.0140

基因表达谱是后基因组时代最先发展起来的高通量技术[1]，其通过测定基因在特定条件下 mRNA 的表达量，能够从整体水平研究基因的结构与功能，揭示特定的生物学过程和疾病发生发展的分子机制，是目前识别和描述与特定表型或者扰动相关的基因表达模式最为有效、准确和高重复性的数据类型，在生物医学领域被广泛应用[2]．基因表达谱的典型应用包括：a．构建可能介导某个重要的生物学过程或者疾病发生发展相关的基因调控网络[3-5]；b．识别与特定表型或者扰动相关的基因表达模式；c．鉴定疾病诊断和预后判断的基因标志物[6]；d．寻找潜在的药靶，用于疾病治疗；e．药物重定位．本文主要介绍基因表达谱分析如何应用于药物发现并最终服务于精准医疗，包括基因表达谱的测定方法、相关数据库和新的数据分析方法．

1 常用基因表达谱测定方法

目前常用的高通量基因表达谱测定方法包括基因芯片、转录组测序以及 L1000 技术．基因芯片概念由 Fodor 等[7]于 1991 年最早提出，其技术原理是利用已知的核酸序列作为探针与互补的靶核酸序列进行杂交，然后使用荧光扫描系统对杂交点的荧光信号进行检测，进而进行定量和定性分析．目前基因芯片广泛用于基因表达谱分析、新基因发现、基因突变和多态性研究，同时在疾病诊断与预测、药物筛选、病原体诊断等方面也发挥重要作用[8]．基因芯片技术虽然得到广泛使用，但是也存在一定局限性，如只能测定已知基因、无法区分相似序列、很难检测低丰度序列、批次效应明显等．

转录组测序是测序技术在转录组研究中的应用[9]，具体指利用高通量测序技术进行测序，全面快速地获取某一物种特定器官或组织在某一状态下的几乎所有转录本．随着测序技术和计算技术的飞速发展，测序时间相比以前大幅缩短、测序成本大幅降低，其用于转录组研究也越来越普遍．相比基因芯片技术，转录组测序技术具有定量准确、可重复性高、检测范围广、敏感性和特异性高等优点，弥补了基因芯片的不足[10]．虽然优势明显，但是目前仍没有取代基因芯片，原因在于目前基因芯片价格仍相对便宜，且芯片数据分析相对简单成熟[11]．科学研究中可以结合两种技术来产生更加可靠的结果，例如可以先使用基因芯片快速筛查大量样本，

* 国家自然科学基金重点项目(U1435222)资助.
** 通讯联系人.
Tel: 010-66932251, 010-66931242, 010-66932211
E-mail: huibai13@hotmail.com, boxc@bmi.ac.cn, sqwang@bmi.ac.cn
收稿日期：2016-04-21，接受日期：2016-09-06

然后用实验结果指导转录组测序，同时也可以使用基因芯片来验证转录组测序的数据.

L1000 技术是整合网络细胞印记库项目(Library of Integrated Network-Based Cellular Signatures，LINCS)中使用的低成本基因表达谱测定技术．其考虑基因表达之间的相关性，将需要测量的基因数目大幅减小，从而达到控制成本的目的．L1000 技术是挑选 978 个标志基因进行测量，进一步通过构建模型外推出其他基因的表达量．实验数据表明这些标志基因的表达可以代表全基因组 80%的信息．虽然存在一定误差，但是该技术可以将成本缩减到全基因组基因芯片成本的 1%，使得测量大规模基因表达谱得以实现.

2 药物发现相关基因表达谱数据库

随着基因芯片技术和转录组测序技术的广泛应用，特别是转录组测序产生了大量的实验数据，传统单纯凭借实验来分析数据已经无法满足科研人员的需求，亟需集成数据库来存储、整理、组合不同实验室、技术平台的基因表达谱数据．目前基因表达谱相关数据库主要有两个：NCBI GEO (Gene Expression Omnibus)[12]和 EBI ArrayExpress[13].

GEO 隶属于美国生物信息中心，自 2000 年发展至今，其已经成为最大、最全面的公共基因表达谱数据库．科研人员可以提交、存储和检索多种形式的数据并且免费使用．GEO 中的数据按照测定平台、样本、系列、数据集进行组织．除了数据检索功能，GEO 还提供 GEO2R 和 GEO BLAST 分析功能．GEO2R 通过比较两组或者多组实验条件下的样本，进行基因差异表达分析，该工具可以方便科研人员快速分析得到关注的基因集合进行后续分析和实验验证．GEO BLAST 则为 GEO 中的数据提供 BLAST[14]比对服务.

ArrayExpress 是欧洲生物信息中心于 2003 年建立的基因表达谱数据库．该数据库包含两部分：实验数据集和基因表达图谱[19]．实验数据集是包含了基因表达的功能基因组学实验数据库，提供查询和下载功能．基因表达图谱数据是一个基于实验数据集加工的子数据集，其提供不同生物条件下的基因表达模式，包括基准表达图谱(Baseline Atlas)和差异表达图谱(Differential Atlas).

除了 GEO 和 ArrayExpress，还存在特定物种或者专项研究的基因表达谱数据库．表 1 列出了基因表达谱相关的数据库．GXD (Gene Expression Database)[15]数据库存储实验小鼠基因表达谱数据．TCGA(The Cancer Genome Atlas)数据库包含了临床肿瘤样本的基因组、转录组、蛋白质组和临床信息，可以用于肿瘤的致病机制阐释和肿瘤药物发现等研究．COSMIC(Catalogue of Somatic Mutations in Cancer)[16]整合了肿瘤相关的组织和细胞系的突变信息，其基因表达谱数据来自 TCGA．CCLE(Cancer Cell Line Encyclopedia)[17]包含 1 000 多种肿瘤细胞

Table 1 Gene expression profile databases for drug discovery

表 1 药物发现相关的基因表达谱数据库

数据库	简介	数据情况	网址
GEO[12]	公共基因表达谱数据库	3 848 个数据集，65 814 个系列，15 487 个芯片平台，1 735 637 个芯片样本	www.ncbi.nlm.nih.gov/geo
ArrayExpress[13]	公共基因表达谱数据库	63 783 次实验，1 921 284 个试验结果	www.ebi.ac.uk/arrayexpress
GXD[15]	实验小鼠基因表达谱	72 824 次表达谱试验，1 509 463 个实验结果	www.informatics.jax.org/expression.shtml
TCGA	临床肿瘤样本转录组	34 种肿瘤类型，每种肿瘤类型包含几十到几百个样本	tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp
COSMIC[16]	肿瘤相关组织和细胞系突变数据库	包含 1 192 776 个样本，9 479 893 个基因突变数据	cancer.sanger.ac.uk/cosmic
CCLE[17]	肿瘤细胞系基因组、转录组	1 046 个肿瘤细胞系的基因表达谱	www.broadinstitute.org/ccle/home
CMap[18]	小分子药物扰动下的细胞系表达谱	1 309 个小分子药物，5 个细胞系，6 100 个基因表达谱	www.broadinstitute.org/cmap
LINCS	基因沉默、基因过表达、小分子化合物扰动下人细胞系的基因表达谱	77 个细胞系，4 372 个基因沉默扰动，3 124 个基因过表达扰动，20 413 个小分子化合物扰动，共 1 328 098 个基因表达谱	www.lincsproject.org

系的基因表达谱数据．关联图谱(Connectivity Map，CMap)数据库测量了 1 309 个小分子药物在 5 个人体细胞系上的表达谱[18]．整合网络细胞印记库项目(Library of Integrated Network-based Cellular Signatures，LINCS)在 CMap 的基础上，进一步扩大细胞系规模和扰动规模，测量人细胞系在基因沉默、基因过表达、小分子化合物作用下的基因表达谱数据．以上丰富的基因表达谱数据库为基于表达谱的药物发现提供了数据基础．

3　基于基因表达谱关联图谱的药物发现

生命医学研究的一个重要挑战在于建立疾病、生理过程和小分子药物之间的联系，因而建立一个数据库来描述不同生物状态对于这个问题的解决至关重要．2006 年 Lamb 等[18]测定 5 个人类癌症细胞系上 1 309 种药物作用后的全基因组表达谱，并利用此数据构建了关联图谱(connectivity map，CMap)数据集．CMap 采用基因集富集分析方法(gene set enrichment analysis，GSEA)[20]，通过与组织细胞在不同生理、病理、药物作用等条件下的基因表达谱印记(signature)进行对比分析，建立基因 - 疾病 - 药物之间的联系．该方法由三部分组成：a．参考基因表达谱数据库；b．查询表达谱印记；c．模式匹配算法．图 1 为 CMap 方法的原理图．

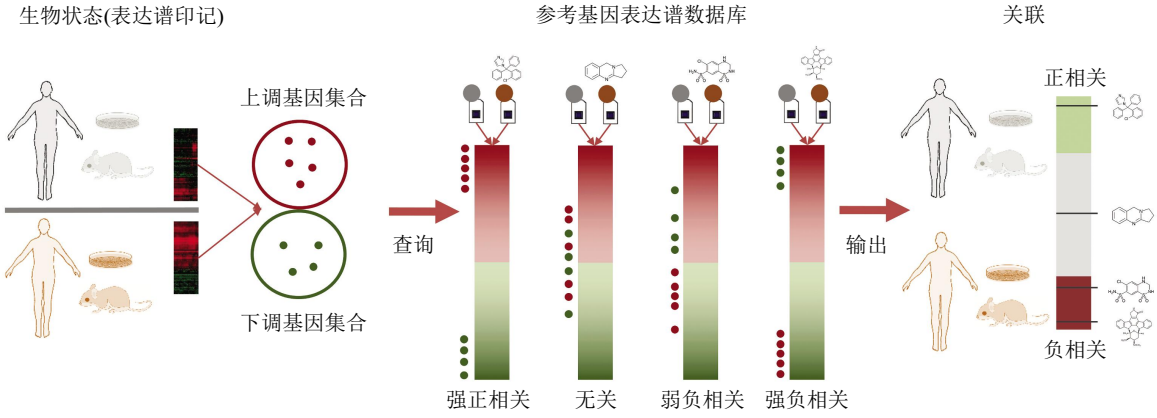

Fig. 1　The connectivity map concept[18]

图 1　关联图谱原理图

大量扰动的人细胞系表达谱构成参考数据库，基因表达谱印记代表诱导产生的目标细胞状态(左)；采用模式匹配算法对查询的表达谱印记进行打分，评估富集程度和方向(中)；根据关联打分排列扰动药物，排名置顶(正相关)和置底(负相关)的药物与查询的目标细胞状态表达谱印记建立联系(右)．

参考基因表达谱数据库包含不同生物状态下的基因表达谱，如不同的扰动类型、细胞系、作用时间、浓度条件等．CMap 第二版数据测量了 1 309 种小分子药物在 5 个人体细胞系上产生的 6 100 个基因表达谱数据，该数据规模在一定程度上限制其潜在应用．2010 年启动的整合网络细胞印记库项目 (Library of Integrated Network-Based Cellular Signatures，LINCS)则在 CMap 的基础上，大大增加了数据规模，旨在在更多不同细胞系上测量不同扰动类型引起的转录组变化，获得更大规模的表达谱数据进行关联网络分析，推动新的生物医学发现．

查询表达谱印记是由用户提供的基因集，用于描述特定的生物状态．根据研究目的的不同，其可以来自人体样本、疾病动物模型、小分子药物作用的细胞和组织等，一般通过相关表达谱数据进行差异表达分析得到．目前存在一些整理的表达谱印记集可用于 CMap 分析．如分子印记数据库(Molecular Signatures Database，MSigDB)[20]，是 GSEA 附带的表达谱印记数据集，最新 MSigDB v5.1 数据集分为 8 个大类，包括标志基因、染色体分类基因、通路相关基因、计算预测基因集、GO 分类基因、肿瘤相关基因等．基因印记数据库(Gene Signature DataBase，GeneSigDB)[21]是一个文献整理得到的表达谱印记数据库，最新第四版数据包含从 1 604 篇文章整理得到的 3 515 个表达谱印

记．这些数据集可以提供不同生物状态下的参考表达谱印记，为相关领域的科学研究提供极大便利．

经过近十年的发展，以 CMap 为先导的表达谱印记比对及关联网络分析方法在基础和临床、系统生物学和药物发现之间架起了桥梁，尤其在抗癌药物发现领域得到了诸多成功应用[22]，如药物重定位、药物先导物发现、药物作用方式阐释等．Wang 等[23]根据基因芯片表达谱数据得到与肺腺癌高度相关的表达谱印记，然后采用该方法发现 HSP90 抑制剂、PPAR 拮抗剂和 PI3K 抑制剂这几类药物与肺腺癌特异性表达谱印记高度负相关，并且实验证实 17-AAG(一种 HSP90 抑制剂)能够抑制肺腺癌细胞生长．Siavelis 等[24]整合 5 个阿尔茨海默病相关的表达谱数据，通过 CMap、SPIEDw[25]、sscMAP[26]、LINCS-L1000 4 个分析工具分别对药物作用细胞表达谱数据集进行分析，发现了 27 个阿尔茨海默病的潜在治疗药物，进一步的信号通路和蛋白质相互作用网络分析，发现这些药物可能是通过影响表皮生长因子受体相关的通路来发挥作用的．表 2 是 2015 年至今基于关联图谱方法的抗癌药物重定位研究．

Table 2 Applications of CMap in anticancer drug repurposing (since 2015)

表 2 关联图谱在抗癌药物重定位中的应用(2015 年至今)

疾病	药物	主要发现	文献
肝癌	Cisplatin，Sorafenib，5-fluorouracil	3 种药物增加组蛋白乙酰化，将生存率低的病人表达谱逆转到生存率高的病人的表达谱.	[27]
胶质母细胞瘤	Pyrvinium	靶向 CD133，体内外实验证实其抑制胶质母细胞瘤自我更新和增殖.	[28]
结直肠癌	氯丙嗪(Chlorpromazine)	抗精神病药物氯丙嗪可以抑制 p53 突变结肠癌细胞系生长，促进其凋亡，SIRT1 为其潜在靶点.	[29]
急性淋巴细胞白血病	类视黄醇	类视黄醇诱导 IKZF1 表达，抑制 BCR-ABL1 型急性淋巴白血病细胞增殖，同时可以增强达沙替尼药效.	[30]
结直肠癌	伊立替康(Irinotecan)等 10 多种药物	采用多个结肠癌数据集构建表达谱印记，筛选出 10 多种药物与表达谱印记关联，包括已知的化疗药物.	[31]
髓母细胞瘤	Alsterpaullone	抑制细胞周期相关基因，提高 3 型髓母细胞瘤生存率.	[32]
急性淋巴细胞白血病	SB225002	SB225002 通过激活 GLIPR1，诱导 B 型和 T 型急性淋巴细胞白血病细胞周期停滞和细胞凋亡.	[33]
弥漫性大 B 细胞淋巴瘤	多西环素(Doxycycline)	抗生素多西环素可以影响多个淋巴瘤形成信号通路，抑制肿瘤细胞生长，CSN5 为其潜在药靶.	[34]
肺腺癌	苯扎贝特(Bezafibrate)	苯扎贝特通过靶向 CDK2，抑制肺腺癌细胞生长.	[35]
肺腺癌	丙戊酸(Valproic acid)	与厄洛替尼联用可以引起酪氨酸激酶抑制剂耐药性肺腺癌细胞凋亡，可能作用于 MAPK 与 AKT 信号通路.	[36]
肺癌	佛司可林(Forskolin)与 6 种 PGE-2 类似物	CRTC1 激活剂与 LKB1 缺失的肺癌相关，COX-2 抑制剂可以抑制 LKB1 缺失的肺腺癌细胞生长.	[37]

随着多组学数据的积累和信息技术尤其是网络模型在药物发现中的应用，传统的“单药物 - 单靶点 - 单疾病”的科研方式已经转向多对多的网络药理学(network pharmacology)[38-42]．通过关联药物、靶点、疾病、病原体等元素，采用网络模型的方法，可以发现大量潜在的假设．而 CMap 很适合构建这样的网络来进行药物重定位和潜在药靶发现研究等．如 Iorio 等[43]构建的药物 - 靶点网络，Hu 等[44]构建的药物 - 疾病网络．Huang 等[45]扩展了 CMap 的数据规模，开发了 DMAP 数据库，构建药物 - 蛋白质关联网络，可服务于药物重定位相关研究．除了用于分析转录组学数据，CMap 也适用于其他组学数据如毒理基因组学(toxicogenomics)[46]、转录因子相互作用数据[47]等．

4 基于基因表达谱重建调控网络的药物发现

对于基因表达谱数据，虽然传统分析方法如聚类[48]、主成分分析[49]、基于线性模型的差异表达分

析[50]可以在一定程度上应用于药物靶点发现、肿瘤分类等方向，但是却无法获取基因之间的大规模调控关系．基因调控网络(gene regulatory networks，GRNs)的出现就旨在构建这些调控关系，而基因表达谱数据的快速积累以及相关分析方法的快速发展，为基因调控网络的广泛应用奠定了基础．基因调控网络本质是建立基因 - 基因之间的调控关系，而这种调控可能通过多种方式实现，如基因 A 编码的转录因子促进基因 B 的表达，或者基因 A 编码的蛋白通过一系列的生物学过程(蛋白质通路或者代谢物)影响 B 的表达．

目前存在大量研究用于构建基因调控网络[51−55]．网络构建方法涉及的算法、构建的网络类型、采用的技术、数据输入输出格式都各有特点．如基于互信息的 ARACNE[3]、CLR[56]等；基于贝叶斯推断的 BANJO[57]、SiGN-BN[58]、PriorPC[59]等；基于常微分方程的 NIR[60]、MIKANA[61]等；基于嵌套效应模型(nested effects models, NEMs)[62]的 HIS[63]；基于稀疏线性回归原理[64]；采用多层分析策略的 GENIMS[65]．除此之外，一些最新的机器学习技术如神经网络也可以用于基因调控网络的构建[66]．目前不存在最优的基因调控网络构建方法，但是可以采用组合方法(ensemble methods)[67]来提高网络的准确性．

基因调控网络在药物发现中的应用主要包括关键药靶的发现，以及药物作用方式的系统阐释，其可以用于发现、诊断、预测和预后的生物标记物[6, 68−70]．利用基因调控网络发现药靶对于复杂疾病如癌症更加有效，因为癌症往往表现为相关通路的基因突变[71]，而这些基因存在直接的相互作用．而要将基因调控网络用于药物设计，则需要整合其他的数据类型，如蛋白质相互作用网络、代谢网络等[72]．通过进一步整合临床数据如生存数据和药物反应数据等，基因调控网络可以辅助服务于个性化医疗[73]．

从基因调控网络发现药靶一般都是采用网络分析方法分析网络的拓扑性质，当然也可以结合其他数据来深入挖掘网络背后的生物学意义，这里介绍两种方法：MARINa (master regulator inference algorithm) [4] 和 DIGGIT (driver-gene inference by genetic-genomic information theory)[5]．

MARINa 用于寻找控制表型过渡的关键转录因子，如正常组织向病理组织过渡的关键转录因子．由于转录组数据难以直接用于预测转录因子调控活性，MARINa 通过分析转录因子的调节子(其激活或者抑制的靶点基因)的状态来预测其活性．采用 MARINa 进行分析依赖一个调控模型和一个基因表达谱印记．调控模型分别采用 ARACNe[3] 与 MINDy[74]来构建转录与翻译后修饰相互作用关系，并结合实验数据、数据库数据、文献挖掘数据对构建的调控模型进行修正．基因表达谱印记通过对疾病与正常状态的表达谱进行差异表达分析得到．MARINa 算法通过定义正负调节子、计算转录因子富集、阴影分析与计算转录因子协同性 4 步，得到主调节因子集合．Lefebvre 等[4]通过构建人 B 细胞的调控模型，采用 MARINa 分析发现 MYB 与 FOXM1 是生发中心(germinal center，GC)增殖的主调节因子，二者协同调控的 80%的基因在生发中心中被激活．

在 MARINa 基础上，DIGGIT 进一步挖掘导致疾病出现的核心基因突变，其基本原理是：任何导致疾病(如癌症)出现的核心突变，一定处于主调节因子的上游，通过影响主调节因子进而改变下游的基因表达．

DIGGIT 分析依赖于大数据量的基因表达谱数据和基因突变数据．其分析分为 5 步：a．MR 分析．采用 MARINa 分析得到主调节因子集合．b．F-CNVG 分析．基于互信息和差异表达分析，去除不重要的拷贝数突变基因．c．MINDy 分析．MINDy 通过分析前面得到的主调节因子和 F-CNVG，结合基因表达谱数据，得到具有调控主调节因子效果的 F-CNVG，缩小 F-CNVG 规模．d．aQTL(expression quantitative trait loci)[75]分析．寻找可以预测主调节因子活性的 F-CNVG．e．条件关联分析．去除因为位置引起的基因突变的关联，进一步筛选保留关键的 F-CNVG．

通过以上步骤，DIGGIT 可以获取到候选核心突变集合．Chen 等[5]采用 DIGGIT 分析发现 KLHL9 缺失可导致间质亚型的胶质母细胞瘤(图 2)，进一步实验发现恢复 KLHL9 表达可抑制肿瘤细胞生长．除了胶质母细胞瘤，DIGGIT 也成功应用于发现乳腺癌和阿尔茨海默病的关键基因突变[5]．DIGGIT 不仅能够得到调控主调节因子的核心突变基因，而且可以解释该突变是如何影响相关基因并最终导致疾病的．所以其不仅可以发掘潜在的药靶，而且还可以解释作用机制．因此结合基因调控网络与其他生物学数据可以更深入解释网络背后的生物学机制，用于药物研发和疾病致病机制阐释．

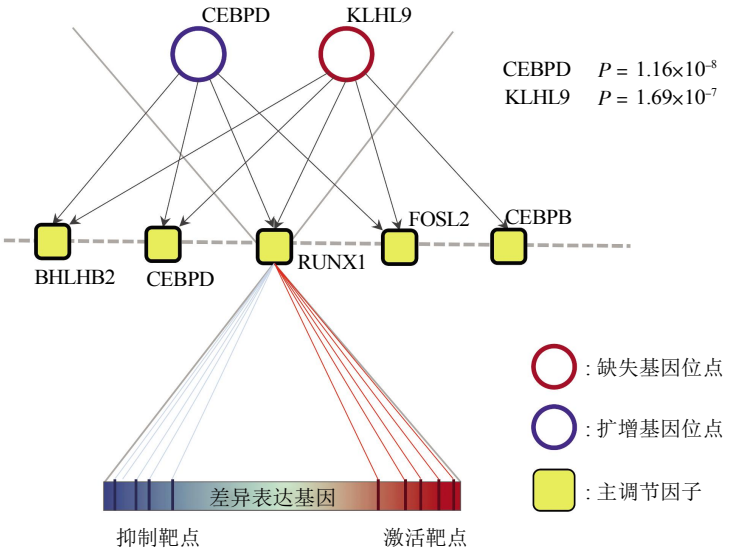

Fig. 2　DIGGIT integrative analysis infers candidate MES-GBM driver mutations[5]

图 2　DIGGIT 整合分析推断间质型胶质母细胞瘤潜在核心突变[5]

DIGGIT 分析发现，间质型胶质母细胞瘤的关键转录因子，上游通路中的 CEBPδ 扩增与 KLHL9 缺失为该肿瘤核心突变，其中关键转录因子由 MARINa 分析得到.

5　基因表达谱与其他组学的融合分析及应用

随着多种组学技术的快速发展，不同层次的生物组学数据的获取及分析方法逐渐完善成熟，与癌症相关的各种组学数据(基因组拷贝数变异、基因突变、甲基化数据以及转录组、蛋白质组、代谢组数据)正在快速积累．通过整合多组学数据，系统地分析疾病发生发展机制，确定最佳药物靶点是精准医学的关键要求和重要发展方向，将对疾病的诊断、个性化治疗和临床用药提供指导.

多组学数据整合涉及大数据处理，现有的大数据处理技术基本是基于统计、机器学习或者网络的方法[76]．其中机器学习方法的优点在于其可以整合大规模异构的生物医学数据[77]，因此在多组学数据整合分析中被广泛使用．机器学习方法可以分为 3 类：

a．有监督学习，如分类和回归问题．对有分类标记的样本进行学习，得到训练模型，然后将新输入的样本映射到分类标记．例如利用疾病和对照样本进行训练，得到模型，输入新的样本判断其是疾病状态还是正常状态．常用的有监督学习技术包括支持向量机(support vector machine，SVM)[78]、核方法(Kernel-based method)[79]、逻辑斯蒂回归(logistic regression)[80].

b．无监督学习，如聚类和降维问题．对无分类标记的样本进行学习，发现数据的结构性知识，得到训练模型，将样本划分到子类．其适应于病人分型以及寻找基因表达谱潜在模式等应用．常用的无监督学习技术如层次聚类(hierarchical clustering)[81]、*K* 均值聚类(*K*-means)[81]、矩阵分解[82]等.

c．半监督学习．结合有监督学习和无监督学习，同时对有分类标记和无分类标记(有分类标记样本较少)的样本进行学习，发现数据的结构，并对无分类标记数据进行预测．例如，为了关联药物 - 疾病，利用半监督学习技术可以先学习已知的药物 - 疾病关联，预测新的药物 - 疾病关系．半监督学习适合用于整合组学数据，常见的技术如正则化网络矩阵分解[83].

5.1　疾病分型

疾病分型指的是基于基因组、转录组、表观基因组和临床数据将病人划分到不同的组，用于预后的判断和治疗决策．目前，疾病分型已经成功用于包括帕金森病、心血管疾病、自身免疫疾病、癌症等疾病的治疗中[84]．癌症是分型应用最多的一种疾病，其不同的分子亚型往往具有显著不同的临床表现．利用数据整合的方法已经发现了一系列癌症类

型如结肠癌[85]、乳腺癌[86]、卵巢癌[87]的亚型．疾病分型在疾病治疗和药物研发中具有重要意义，如可以针对特定亚型开发治疗药物，提高药物研发效率.

基因表达谱是疾病分型中使用最广泛的数据类型，采用无监督聚类算法如层次聚类[88]、K 均值聚类[89]、非负矩阵分解[90]可以根据基因表达量将样本划分到不同的子类．Wang 等[88]采用 TCGA 的乳腺癌和肺癌基因表达谱数据，基于差异表达基因对样本进行层次聚类，将乳腺癌和肺癌分别划分为 4 个和 5 个亚型，亚型之间具有显著不同的生存率．并且结合基因组突变数据将每一个亚型中突变频率最高的 15 个基因映射到蛋白质相互作用网络，寻找最优连接子网，这些子网构成核心突变模型，可以作为药物重定位的潜在靶点．Vaske 等[91]开发的 PARADIGM 可以预测病人样本的异常通路，从而用于疾病分型．PARADIGM 应用于 TGCA 的多形性胶质母细胞瘤基因表达谱数据和拷贝数突变数据，将样本分为 4 个生存率差异显著的亚型．

最新的方法则通过整合更多的数据类型，如拷贝数突变、甲基化数据、转录组数据、蛋白组数据和分子相互作用数据，来进行更加准确的分型．Shen 等[92]开发了 iCluster 工具包用于聚类、数据融合、特征选择、降维等一系列分析，被广泛应用于各种癌症的分型．Curtis 等[93]利用 iCluster 对 2000 个乳腺癌样本进行分类，得到 10 个子类，并且发现基因组突变数据和转录组数据之间的强烈关联．Shen 等[94]应用 iCluster 对 TCGA 多形性胶质母细胞瘤数据进行分析，通过整合拷贝数突变、甲基化数据和基因表达谱数据，发现了 3 个显著的亚型，与之前单一通过基因表达谱数据预测的结果形成显著对比．

为了解决聚类方法对基因 - 病人高维度矩阵可扩展性差的问题，Wang 等[95]提出了相似性网络融合方法(similarity network fusion，SNF)，该方法结合转录组数据、甲基化数据和 microRNA表达谱数据，构建病人 - 病人相似性矩阵．其原理如下：首先，对于每一种数据类型，构建病人 - 病人的加权网络，节点代表病人，边代表病人之间的相似度．然后，SNF 考虑所有数据类型标准化边的权重．最后，SNF 采用信息扩散的方法将不同数据类型构建的网络融合成一个网络，采用谱聚类算法将融合网络划分成不同的类(图 3)．由于 SNF 构建的是病人 - 病人矩阵，矩阵规模远小于之前方法采用的基因 - 病人矩阵，因此其收敛速度快、扩展性好．应用 SNF 分析 TCGA 5 种癌症数据的结果显示，SNF 显著优于基于单一数据类型的分析方法，适用于癌症分型和生存期预测[95].

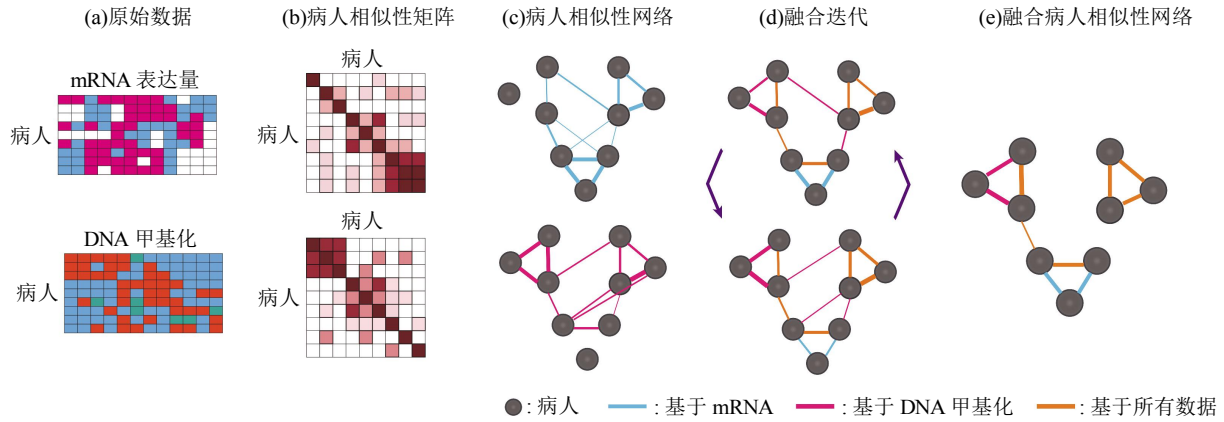

Fig. 3　Illustrative example of SNF steps[95]

图 3　SNF 计算过程[95]

(a) 来自相同病人样本的转录组与甲基化数据．(b) 不同数据类型构建的相似性矩阵．(c) 病人 - 病人相似性网络，节点代表病人，边代表病人之间的相似性．(d) SNF 采用信息扩散更新网络权重．(e) 最终生成的融合网络，不同颜色代表相应数据类型．

目前采用多组学数据进行分型多是基于基因表达谱和拷贝数突变数据，由于数据本身存在噪声，预测的亚型并不能很好地和临床数据、患者生存率数据吻合[85]．因此，Hofree 等[96]采用体细胞突变数据进行癌症分型，提出了网络分层方法(network-based stratification，NBS)．由于癌症的异质性，相同癌症的不同病人其突变往往不同，单纯从单基因角度进行分型难度较大．但是有研究表明，尽管不

同的癌症样本之间的突变差异较大，其涉及的通路往往很相似[97]，与肿瘤相关的核心基因突变基本都是出现在细胞生存、细胞命运和基因组维护相关的通路中．NBS 整合了体细胞突变数据与通路数据，其基于非负矩阵分解，采用半监督学习方法进行聚类，在聚类中采用分子网络作为先验知识，不仅考虑了体细胞突变之间的相似性，也考虑了分子网络中突变基因的近邻程度．NBS 应用于 TCGA 的卵巢癌、子宫癌和肺癌数据，取得了很好的预测结果.

由于矩阵分解方法十分适合用于整合异构的数据，Gligorijevic 等[98]扩展 NBS 的方法，进一步整合药物数据，可以同时进行癌症分型、药物重定位和生物标志物发现．

以上列举了应用多组学数据进行疾病分型的各种算法，但是需要注意这些方法存在一定局限性．疾病分型没有标准答案，在临床数据上得到的生存率显著差异的亚型并不能确保其分型结果是最合理的．而且，这些方法都不能整合各种数据类型，如 SNF 只能处理连续变量，而对体细胞突变这种离散数据则无法处理．事实上，很少有研究来整合离散的体细胞突变数据和连续的基因表达谱、甲基化数据[99].

5.2 药物重定位

目前已经存在大量用于药物重定位的计算方法[100]．Dudley 等[101]建议从数据角度将这些方法分成基于药物的方法和基于疾病的方法．a．基于药物的方法采用某种指标度量药物之间的相似度，如化学相似性[102]、药物作用后的基因表达谱相似性[18]、副作用相似性[103]．然后根据相似性对药物进行聚类，如果一个新药和某个组内的药物表现相同的反应，则可以推测该药具有与该组中药物相似的作用，用于药物重定位．b．基于疾病的方法度量疾病之间的相似度，如表型相似性[104]、疾病症状相似性[105]，然后对疾病进行聚类，通过结合组内已知的疾病 - 药物关系来推测新的疾病 - 药物关系．c．其他的一些方法基于靶点的相似度[106]来预测新药物，如基于蛋白质序列相似度[107]、三维结构的相似度[108]．以上的三种方法都是基于相似度的药物重定位方法，采用的技术一般是机器学习或者基于网络的方法．当然也存在一些其他的药物重定位方法如分子对接(molecular docking)，是计算机辅助药物设计领域的重要技术．分子对接依赖蛋白质结构信息和对接算法，虽然目前蛋白质结构数据库(Protein Data Bank，PDB)[109]中存储的蛋白质结构信息不断增加，但是仍然有限．并且与基于转录组的药物发现相比，分子对接方法中涉及的受体柔性、计算量较大、不同对接算法打分函数设计等问题都在一定程度上限制了其广泛应用[110].

上面列举了各种基于相似度的药物重定位方法，这些方法都能在一定程度上为药物发现提供指导．实际上，我们可以整合多组学数据，综合多种相似度指标来进一步提高药物重定位结果的准确度．利用多组学数据进行药物重定位，需要整合药物、化学、基因和临床等数据．而综合考虑病人的分子多样性，用于疾病的诊断和给药是精准医疗的最终目标[111]．因此，整合多组学数据用于药物重定位和个性化医疗越来越受到重视．

Napolitano 等[112]采用核方法整合了药物化学相似性、蛋白质 - 蛋白质相互作用数据、药物作用的基因表达谱数据，然后采用支持向量机(SVM)对药物进行分类，经过训练的 SVM 在已知药物分类上得到了 78%的精度，而得分最高分类错误的药物则用于药物重定位．

Gottlieb 等[113]开发了 PREDICT (PREdicting Drug IndiCaTions)，其首先分别采用 5种和 6 种数据计算药物 - 药物相似性和疾病 - 疾病相似性．然后根据这些相似性矩阵，计算药物 - 疾病相似性．最后，根据已知药物 - 疾病关联，PREDICT 训练逻辑斯蒂回归分类器，预测药物的新适应症．PREDICT 预测精度很高，其 AUC[114]可以达到 0.92，可以作为一个通用框架，进一步整合基因表达谱数据，用于个性化医疗．

目前整合数据进行药物重定位基本采用的都是网络关联的方法，而其中整合的多组学数据均是用来度量药物、疾病或者基因之间的相似性，基因表达谱在其中的作用并不显著．而疾病分型和生物标记物发现则基本都依赖转录组数据，虽然这些更加接近临床，但是其结果对于药物的研发至关重要．特别是在癌症药物研发领域，如何结合癌症的分子特征，开发针对性的靶向药物，对于提高病人的预后尤为关键．

6 结　　语

作为一种有效的技术方法，基因表达谱在药物的早期发现、后期研发以及临床用药指导都发挥着重要作用[115]，特别是在肿瘤学中根据肿瘤亚型指导临床治疗中成效显著．随着测序技术的发展，转录组测序成本不断降低，将使得在成本可控的条件

下测量大量不同条件下的基因表达谱数据可行，结合最新的基因表达谱分析方法，将为药物发现提供更为广阔的空间.

本小组曾将基因表达谱关联图谱的方法分别应用于抗感染药物、抗癌药物和抗阿尔茨海默病药物重定位研究，并且初步实验验证了若干潜在的治疗药物．如在抗癌药物研究中，我们基于 LINCS 小分子化合物扰动细胞反应表达谱数据、整合特定临床患者癌症基因组数据和药物敏感性基因组数据，系统评价了 LINCS 小分子化合物单独或联合使用时对 6 种不同类型癌症的修正潜能并预测药物活性及特异性，并且通过体内外实验多指标评价药物(组合)的癌症修正效果．由于关联图谱预测的药物依赖于参考数据库中的小分子化合物，因此，进一步扩大小分子化合物规模，将为基于表达谱的高通量药物筛选提供可能．此外，LINCS 中不同扰动类型的数据(基因、配体、小分子化合物)也可用于寻找相应药物的作用靶点.

与此同时，基于基因表达谱重建调控网络可以在很大程度上辅助用于药物研究，发现决定疾病(如癌症)的关键突变和调控因子，作为潜在的药物作用靶点．随着多组学时代的来临，融合多组学数据是未来发展趋势，而其应用不仅仅局限于疾病分型和药物重定位方向．结合最新的人工智能技术[116]，其已经服务于精准医疗，如国际商业机器公司(IBM)的沃森(Waston)系统，可以将患者的组学数据与构建的疾病知识库比对，根据患者的遗传学特征提供最优的治疗方案．而组学数据的进一步增长，将进一步提高该系统的准确度．目前，美国和加拿大的多家肿瘤中心已经部署沃森系统，并将其应用于癌症诊断与治疗中.

当然，基因表达谱分析也存在一定不足，如样本少、特征太多引起的统计检验有效性问题．此外，如何整合多组学异构数据，如何将基因表达谱印记转化到相关的临床应用也是目前表达谱应用面临的重大挑战．相信随着转录组测序技术和相关数据分析技术的持续发展，未来基因表达谱在药物研发中将发挥更加重要的作用.

参 考 文 献

[1] Lockhart D J, Winzeler E A. Genomics, gene expression and DNA arrays. Nature, 2000, **405**(6788): 827−836

[2] Gligorijevic V, Malod-Dognin N, Przulj N. Integrative methods for analysing big data in precision medicine. Proteomics, 2015

[3] Margolin A A, Nemenman I, Basso K, *et al*. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 2006, **7**(Suppl 1): S7

[4] Lefebvre C, Rajbhandari P, Alvarez M J, *et al*. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Molecular Systems Biology, 2010, **6**(1): 377

[5] Chen J C, Alvarez M J, Talos F, *et al*. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. Cell, 2014, **159**(2): 402−414

[6] Dehmer M, Mueller L A, Emmert-Streib F. Quantitative network measures as biomarkers for classifying prostate cancer disease states: a systems approach to diagnostic biomarkers. PloS One, 2013, **8**(11): e77602

[7] Fodor S P, Read J L, Pirrung M C, *et al*. Light-directed, spatially addressable parallel chemical synthesis. Science, 1991, **251**(4995): 767−773

[8] 张 骞, 盛 军. 基因芯片技术的发展和应用. 中国医学科学院学报, 2008, **03**: 344−347
Zhang Q, Sheng J. Acta Academiae Medicinae Sinicae, 2008, **03**: 344−347

[9] 王兴春, 杨致荣, 王 敏, 等. 高通量测序技术及其应用. 中国生物工程杂志, 2012, **01**: 109−114
Wang X C, Yang Z R, Wang M, *et al*. China Biotechnology, 2012, **01**: 109−114

[10] Hurd P J, Nelson C J. Advantages of next-generation sequencing versus the microarray in epigenetic research. Briefings in Functional Genomics & Proteomics, 2009, **8**(3): 174−183

[11] Zhao S, Fung-Leung W P, Bittner A, *et al*. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS One, 2014, **9**(1): e78644

[12] Barrett T, Wilhite S E, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Research, 2013, **41** (Database issue): D991−995

[13] Kolesnikov N, Hastings E, Keays M, *et al*. ArrayExpress update-simplifying data submissions. Nucleic Acids Research, 2015, **43**(D1): D1113−D1116

[14] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool. Journal of Molecular Biology, 1990, **215**(3): 403−410

[15] Finger J H, Smith C M, Hayamizu T F, *et al*. The mouse gene expression database: New features and how to use them effectively. Genesis, 2015, **53**(8): 510−522

[16] Forbes S A, Beare D, Gunasekaran P, *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research, 2015, **43**(Database issue): D805−811

[17] Barretina J, Caponigro G, Stransky N, *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012, **483**(7391): 603−607

[18] Lamb J, Crawford E D, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006, **313**(5795): 1929−1935

[19] Petryszak R, Keays M, Tang Y A, *et al*. Expression Atlas update-an

integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Research, 2016, **44** (D1): D746-752

[20] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102**(43): 15545-15550

[21] Culhane A C, Schroder M S, Sultana R, *et al*. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Research, 2012, **40**(Database issue): D1060-1066

[22] Qu X A, Rajpal D K. Applications of Connectivity Map in drug discovery and development. Drug Discovery Today, 2012, **17**(23-24): 1289-1298

[23] Wang G, Ye Y, Yang X, *et al*. Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. PloS One, 2011, **6**(1): e14573

[24] Siavelis J C, Bourdakou M M, Athanasiadis E I, *et al*. Bioinformatics methods in drug repurposing for Alzheimer's disease. Briefings in Bioinformatics, 2015

[25] Williams G. SPIEDw: a searchable platform-independent expression database web tool. BMC Genomics, 2013, **14**: 765

[26] Zhang S D, Gant T W. sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. BMC Bioinformatics, 2009, **10**: 236

[27] Gillet J P, Andersen J B, Madigan J P, *et al*. A gene expression signature associated with overall survival in patients with hepatocellular carcinoma suggests a new treatment strategy. Molecular Pharmacology, 2016, **89**(2): 263-272

[28] Venugopal C, Hallett R, Vora P, *et al*. Pyrvinium targets CD133 in human glioblastoma brain tumor-initiating cells. Clin Cancer Res, 2015, **21**(23): 5324-5337

[29] Lee W Y, Lee W T, Cheng C H, *et al*. Repositioning antipsychotic chlorpromazine for treating colorectal cancer by inhibiting sirtuin 1. Oncotarget, 2015, **6**(29): 27580-27595

[30] Churchman M L, Low J, Qu C X, *et al*. Efficacy of retinoids in IKZF1-mutated BCR-ABL1 acute lymphoblastic leukemia. Cancer Cell, 2015, **28**(3): 343-356

[31] Wen Q, O'reilly P, Dunne P D, *et al*. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. BMC Systems Biology, 2015, **9**(Suppl 5): S4

[32] Faria C C, Agnihotri S, Mack S C, *et al*. Identification of alsterpaullone as a novel small molecule inhibitor to target group 3 medulloblastoma. Oncotarget, 2015, **6**(25): 21718-21729

[33] De Vasconcellos J F, Laranjeira A B A, Leal P C, *et al*. SB225002 induces cell death and cell cycle arrest in acute lymphoblastic leukemia cells through the activation of GLIPR1. PloS One, 2015, **10**(8): 19

[34] Pulvino M, Chen L, Oleksyn D, *et al*. Inhibition of COP9-signalosome (CSN) deneddylating activity and tumor growth of diffuse large B-cell lymphomas by doxycycline. Oncotarget, 2015, **6**(17): 14796-14813

[35] Liu X, Yang X, Chen X, *et al*. Expression profiling identifies bezafibrate as potential therapeutic drug for lung adenocarcinoma. Journal of Cancer, 2015, **6**(12): 1214-1221

[36] Zhuo W, Zhang L, Zhu Y, *et al*. Valproic acid, an inhibitor of class I histone deacetylases, reverses acquired Erlotinib-resistance of lung adenocarcinoma cells: a Connectivity Mapping analysis and an experimental study. American Journal of Cancer Research, 2015, **5**(7): 2202-2211

[37] Cao C, Gao R, Zhang M, *et al*. Role of LKB1-CRTC1 on glycosylated COX-2 and response to COX-2 inhibition in lung cancer. Journal of the National Cancer Institute, 2015, **107**(1): 358

[38] 潘家祜. 基于网络药理学的药物研发新模式. 中国新药与临床杂志, 2009, **10**: 721-726
Pan J H. Chinese Journal of New Drugs and Clinical Remedies, 2009, **10**: 721-726

[39] 刘艾林, 杜冠华. 网络药理学:药物发现的新思想. 药学学报, 2010, **12**: 1472-1477
Liu A L, Du G H. Acta Pharmaceutica Sinica, 2010, **12**: 1472-1477

[40] 周文霞, 程肖蕊, 张永祥. 网络药理学: 认识药物及发现药物的新理念. 中国药理学与毒理学杂志, 2012, **01**: 4-9
Zhou W X, Cheng X R, Zhang Y X. Chinese Journal of Pharmacology and Toxicology, 2012, **01**: 4-9

[41] 刘志华, 孙晓波. 网络药理学:中医药现代化的新机遇. 药学学报, 2012, **06**: 696-703
Liu Z H, Sun X B. Acta Pharmaceutica Sinica, 2012, **06**: 696-703

[42] Hopkins A L. Network pharmacology: the next paradigm in drug discovery. Nature Chemical Biology, 2008, **4**(11): 682-690

[43] Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci USA, 2010, **107**(33): 14621-14626

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

[45] Huang H, Nguyen T, Ibrahim S, *et al*. DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. BMC Bioinformatics, 2015, **16**(Suppl 13): S4

[46] Toyoshiba H, Sawada H, Naeshiro I, *et al*. Similar compounds searching system by using the gene expression microarray database. Toxicology Letters, 2009, **186**(1): 52-57

[47] Lachmann A, Xu H, Krishnan J, *et al*. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics, 2010, **26**(19): 2438-2444

[48] Andreopoulos B, An A, Wang X, *et al*. A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in Bioinformatics, 2009, **10**(3): 297-314

[49] Clarke R, Ressom H W, Wang A, *et al*. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer, 2008, **8**(1): 37-49

[50] Smyth G K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 2004, **3**: Article3

[51] Marbach D, Prill R J, Schaffter T, *et al*. Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci USA, 2010, **107**(14): 6286-6291

[52] Penfold C A, Wild D L. How to infer gene networks from expression profiles, revisited. Interface Focus, 2011, **1**(6): 857-870

[53] Hurley D, Araki H, Tamada Y, *et al*. Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. Nucleic Acids Research, 2012, **40** (6): 2377-2398

[54] Madhamshettiwar P B, Maetschke S R, Davis M J, *et al*. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Medicine, 2012, **4**(5): 41

[55] 雷耀山, 史定华, 王翼飞. 基因调控网络的生物信息学研究. 自然杂志, 2004, **01**: 7-12
Lei Y S, Shi D H, Wang Y F. Bioinformatics research of gene regulatory networks. Nature Magazine, 2004, **01**: 7-12

[56] Faith J J, Hayete B, Thaden J T, *et al*. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biology, 2007, **5**(1): e8

[57] Yu J, Smith V A, Wang P P, *et al*. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics, 2004, **20**(18): 3594-3603

[58] Kim S, Imoto S, Miyano S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Bio Systems, 2004, **75**(1-3): 57-65

[59] Ghanbari M, Lasserre J, Vingron M. Reconstruction of gene networks using prior knowledge. BMC Systems Biology, 2015, **9**: 84

[60] Gardner T S, Di Bernardo D, Lorenz D, *et al*. Inferring genetic networks and identifying compound mode of action *via* expression profiling. Science, 2003, **301**(5629): 102-105

[61] Wildenhain J, Crampin E J. Reconstructing gene regulatory networks: from random to scale-free connectivity. Systems Biology, 2006, **153**(4): 247-256

[62] Markowetz F, Kostka D, Troyanskaya O G, *et al*. Nested effects models for high-dimensional phenotyping screens. Bioinformatics, 2007, **23**(13): i305-312

[63] Snijder B, Liberali P, Frechin M, *et al*. Predicting functional gene interactions with the hierarchical interaction score. Nature Methods, 2013, **10**(11): 1089-1092

[64] Omranian N, Eloundou-Mbebi J M, Mueller-Roeber B, *et al*. Gene regulatory network inference using fused LASSO on multiple data sets. Scientific Reports, 2016, **6**: 20533

[65] Wu J, Zhao X, Lin Z, *et al*. Large scale gene regulatory network inference with a multi-level strategy. Molecular bioSystems, 2016, **12**(2): 588-597

[66] Rubiolo M, Milone D H, Stegmayer G. Mining gene regulatory networks by neural modeling of expression time-series. IEEE/ACM Transactions on Computational Biology and Bioinformatics/ IEEE, ACM, 2015, **12**(6): 1365-1373

[67] Zhang H. S B H. Recursive Partitioning and Applications [M]. New York: NY: Springer, 2010

[68] Chuang H Y, Lee E, Liu Y T, *et al*. Network-based classification of breast cancer metastasis. Molecular Systems Biology, 2007, **3**(1): 140

[69] Ben-Hamo R, Efroni S. Gene expression and network-based analysis reveals a novel role for hsa-miR-9 and drug control over the p38 network in glioblastoma multiforme progression. Genome Medicine, 2011, **3**(11): 77

[70] Chen L, Xuan J, Riggins R B, *et al*. Identifying cancer biomarkers by network-constrained support vector machines. BMC Systems Biology, 2011, **5**(161)

[71] Hanahan D, Weinberg R A. Hallmarks of cancer: the next generation. Cell, 2011, **144**(5): 646-674

[72] Barabasi A L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 2011, **12**(1): 56-68

[73] Chan I S, Ginsburg G S. Personalized medicine: progress and promise. Annual Review of Genomics and Human Genetics, 2011, **12**: 217-244

[74] Wang K, Saito M, Bisikirska B C, *et al*. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nature Biotechnology, 2009, **27**(9): 829-839

[75] Yang X, Deignan J L, Qi H, *et al*. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet, 2009, **41**(4): 415-423

[76] Greene C S, Tan J, Ung M, *et al*. Big data bioinformatics. Journal of Cellular Physiology, 2014, **229**(12): 1896-1900

[77] Gligorijevic V, Przulj N. Methods for biological data integration: perspectives and challenges. Journal of the Royal Society, Interface/ the Royal Society, 2015, **12**(112)

[78] Vapnik V N V, V. Statistical learning theory [M]. New York: Wiley, 1998

[79] Scholkopf B S, A. J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond [M]. Cambridge, MA, USA: MIT Press, 2001

[80] Freedman D A. Statistical models: theory and practice [M]. cambridge university press, 2009

[81] Hartigan J A. Clustering Algorithms [M]. New York, NY, USA: John Wiley & Sons, Inc., 1975

[82] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, **401**(6755): 788-791

[83] Wang F, Li, T. & Zhang, C. Semi-supervised clustering *via* matrix factorization [M]. Proceedings of the SIAM International Conference on Data Mining. Atlanta, Georgia, USA. 2008

[84] Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. Ieee Intell Syst, 2015, **30**(4): 70-75

[85] Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012, **487** (7407): 330-337

[86] Cancer Genome Atlas N. Comprehensive molecular portraits of

human breast tumours. Nature, 2012, **490**(7418): 61-70

[87] Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature, 2011, **474**(7353): 609-615

[88] Wang L, Li F H, Sheng J T, *et al*. A computational method for clinically relevant cancer stratification and driver mutation module discovery using personal genomics profiles. BMC Genomics, 2015, **16**(Suppl 7): S6

[89] De Souto M C, Costa I G, De Araujo D S, *et al*. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics, 2008, **9**: 497

[90] Brunet J P, Tamayo P, Golub T R, *et al*. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci USA, 2004, **101**(12): 4164-4169

[91] Vaske C J, Benz S C, Sanborn J Z, *et al*. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 2010, **26**(12): i237-i245

[92] Shen R, Olshen A B, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 2009, **25**(22): 2906-2912

[93] Curtis C, Shah S P, Chin S F, *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012, **486**(7403): 346-352

[94] Shen R, Mo Q X, Schultz N, *et al*. Integrative subtype discovery in glioblastoma using iCluster. PloS One, 2012, **7**(4)

[95] Wang B, Mezlini A M, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods, 2014, **11**(3): 333-337

[96] Hofree M, Shen J P, Carter H, *et al*. Network-based stratification of tumor mutations. Nature Methods, 2013, **10**(11): 1108-1115

[97] Vogelstein B, Papadopoulos N, Velculescu V E, *et al*. Cancer genome landscapes. Science, 2013, **339**(6127): 1546-1558

[98] Gligorijevic V, Malod-Dognin N, Przulj N. Patient-specific data fusion for cancer stratification and personalised treatment. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2016, **21**(6): 321-332

[99] Mo Q, Wang S, Seshan V E, *et al*. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci USA, 2013, **110**(11): 4245-4250

[100]谢达菲, 李鹏, 李非, 等. 药物重定位的计算分析方法. 生物化学与生物物理进展, 2012, **39**(11): 1029-1036
Xie D F, Li P, Li F, *et al*. Progress in Biochemistry and Biophysics, 2012, **39**(11): 1029-1036

[101]Dudley J T, Deshpande T, Butte A J. Exploiting drug-disease relationships for computational drug repositioning. Briefings in Bioinformatics, 2011, **12**(4): 303-311

[102]Keiser M J, Setola V, Irwin J J, *et al*. Predicting new molecular targets for known drugs. Nature, 2009, **462**(7270): 175-181

[103]Campillos M, Kuhn M, Gavin A C, *et al*. Drug target identification using side-effect similarity. Science, 2008, **321**(5886): 263-266

[104]Van Driel M A, Bruggeman J, Vriend G, *et al*. A text-mining analysis of the human phenome. European Journal of Human genetics : EJHG, 2006, **14**(5): 535-542

[105]Zhou X, Menche J, Barabasi A L, *et al*. Human symptoms-disease network. Nature Communications, 2014, **5**: 4212

[106]Ding H, Takigawa I, Mamitsuka H, *et al*. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Briefings in Bioinformatics, 2014, **15**(5): 734-747

[107]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): i232-240

[108]Minai R, Matsuo Y, Onuki H, *et al*. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. Proteins, 2008, **72**(1): 367-381

[109]Berman H M, Westbrook J, Feng Z, *et al*. The protein data bank. Nucleic Acids Research, 2000, **28**(1): 235-242

[110]Meng X Y, Zhang H X, Mezei M, *et al*. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des, 2011, **7**(2): 146-157

[111]Li Y Y, Jones S J M. Drug repositioning for personalized medicine. Genome Medicine, 2012, **4**(3): 27

[112]Napolitano F, Zhao Y, Moreira V M, *et al*. Drug repositioning: a machine-learning approach through data integration. Journal of Cheminformatics, 2013, **5**(1): 30

[113]Gottlieb A, Stein G Y, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology, 2011, **7**(1): 496

[114]Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. Journal of Evaluation in Clinical Practice, 2006, **12**(2): 132-139

[115]Bai J P, Alekseyenko A V, Statnikov A, *et al*. Strategic applications of gene expression: from drug discovery/development to bedside. The AAPS Journal, 2013, **15**(2): 427-437

[116]李 渊, 骆志刚, 管乃洋, 等. 生物医学数据分析中的深度学习方法应用. 生物化学与生物物理进展, 2016, **43**(5): 472-483
Li Y, Luo Z G, Guan N Y, *et al*. Progress in Biochemistry and Biophysics, 2016, **43**(5): 472-483

Gene Expression Profile Analysis for Drug Discovery and Precision Medicine*

LIU Yang[1)], BAI Hui[2)**], TAO Huan[3)], HE Song[4)], HUANG Xin[4)], BO Xiao-Chen[4)**], WANG Sheng-Qi[4)**]

([1)] *No. 302 Hospital of PLA, Beijing* 100039, China; [2)] *No. 451 Hospital of PLA, Xi'an* 710054, China;

[3)] *The Fourth Military Medical University, Xi'an* 710032, China;

[4)] *Institute of Radiation Medicine, Academy of Military Medical Sciences, Beijing* 100850, China)

Abstract As an important part of functional genomics, gene expression profiling plays an important role in many fields, such as biology, medicine and drug discovery. With the advent of precision medicine, integration of multi-omics data for personalized health care is becoming the trend of future medicine. In this review, Relevant databases of gene expression profile were introduced first. And then three general methods for gene expression profile analysis, *i.e.*, connectivity map method, gene regulatory network method and multi-omics data integration methods were focused. The latest usage of these methods in drug discovery, especially in cancer drug development was reviewed. This review would serve as a guide for transcriptome-based drug discovery.

Key words gene expression profile, connectivity map, library of integrated network-based cellular signatures, gene regulatory network, multi-omics data integration, precision medicine

DOI: 10.16476/j.pibb.2016.0140

* This work was supported by a grant from The National Natural Science Foundation of China (U1435222).

**Corresponding author.

BAI Hui. Tel: 86-10-66932251, E-mail: huibai13@hotmail.com

BO Xiao-Chen. Tel: 86-10-66931242, E-mail: boxc@bmi.ac.cn

WANG Sheng-Qi. Tel: 86-10-66932211, E-mail: sqwang@bmi.ac.cn

Received: April 21, 2016 Accepted: September 6, 2016