·综述·

基于组学数据的生物学模块识别方法研究进展*

刘芝霖[1] 荣志炜[1] 黄吉科[1] 宋佳丽[1] 俞轶培[1] 侯 艳[1,2△]

【中图分类号】 R195.1 【文献标识码】 A DOI 10.11783/j.issn.1002-3674.2024.04.033

组学数据在分子层面上加深了研究者对表型相似疾病的辨别能力及对认识不足疾病的理解,对组学数据的分析能够在疾病的诊断与治疗中发挥至关重要的作用。组学分子间存在复杂的调控关系,过去的研究表明细胞的功能是模块化的[1],在疾病的实际进展中,关键的调控网络出现异常往往表现为网络中所有分子表达水平都发生变化[2],因此表达模式相似的分子很可能存在共调控或功能相关的情况[3]。利用组学数据这一生物学特点,从组学数据中识别出特定功能的生物学模块,有利于疾病相关信号调节机制的研究[4]。识别出模块后,模块内所有分子将被看作整体分析,减少了假阳性与假阴性,以整体作为生物标志物也更加稳健。模块的识别有利于反映不同疾病的细微差异[5],发现新的致病基因、生物标志物[6],将识别的模块融入深度学习等“黑箱”模型,可以有效提升其解释性[7]。生物学模块的识别是组学数据分析中的重要思想,但目前国内还缺少相关的综述,因此本文将填补相关研究空白,介绍模块思想的产生,按原理分类模块识别的方法,并提供方法评估的思路,以促进组学数据分析的发展。

模块思想的产生

在生物学模块思想产生前,传统的组学分析方法从数据中寻找差异基因,由此进行疾病机制研究和生物标志物发现,但从有限样本中发现的差异基因或生物标志物特异度较高,假阳性多,可重复性差[8],甚至与疾病并没有关系[9]。为了解释结果,传统方法一般会对发现的差异基因做生物学功能富集,找到其共同发挥作用的功能模块,解释该模块在疾病发展中的作用,以证明被富集的差异基因的正确性,但该做法难以解释未被富集的差异基因和模块中没有显著差异的基因。没有差异表达的基因也可能在细胞中起到重要作用,由多基因影响的复杂疾病其每个基因单独的影响都很小[10],模块分子间的相互作用能够放大一系列有关联的微小信号[11],从而导致模块整体活性变化,改变细胞的信号程序调控[12],仅基于差异基因的富集难以发现这一类模块。此外,组学分子相互作用网络呈现无标度分布,其典型特征是在网络中的大部分节点只和较少的节点连接,而有极少的关键节点与非常多的节点连接,使其可以划分为不同的功能模块[13],这也与细胞功能的模块化相对应。以上的事实使研究者发现,通过识别差异基因做功能富集的方法存在一定缺陷,若是直接识别具有共同功能的组学分子作为模块,既符合生物学特点,也能够从原理上减少假阳性与假阴性,便于解释生理机制,由此便产生了直接识别模块,以模块为基础进行后续分析的思想。

虽然现有数据库中的基于实验证据的通路数据或功能基因集也可以直接作为模块使用,如京都基因与基因组百科全书[14](kyoto encyclopedia of genes and genomes,KEGG)通路或基因本体[15](gene ontology,GO)生物功能,但因生物学上的认识有限,这些模块覆盖基因较少。疾病的复杂性与异质性使得每个患者的致病机制都可能存在细微的差异,可能对应于不同的治疗方法,患者特异的模块有利于疾病的预测与个性化治疗。使用固定的模块信息无法反映特定的疾病或患者的特异改变,理想的模块应该是在现有通路基础上针对疾病或患者特定进行优化的结果,基于统计方法识别的模块不受已知的通路信息限制,能够依靠数据发现不同细胞类型的表达模式[5]和患者特异的疾病机制[6]。目前,在了解较少的癌症驱动模块或疾病特异模块中,模块的识别是重要的疾病机制探索方法[4],能为数据库的更新提供实验方向。因此,通过统计方法识别模块不会被数据库中现成的模块取代,反而具有很强的研究价值。

模块识别方法分类

生物学模块的核心思想为识别一组参与共同功能的基因作为整体,以此减少假阳性与假阴性,增强结果的解释性[16]。不同的研究在应用该思想时可能因为不同的研究目的将模块具体化为子网络、通路、基因集、调控复合体等生物学实体,而本综述将所有采用这一思想的文章都纳入模块识别方法。因此,模块识别

*基金项目:科技部国家重点研发计划(2021YFF0901401);国家自然科学基金(82173615;82204158);中央高校基本科研业务费专项资金

1.北京大学公共卫生学院生物统计系(100191)
2.北京大学临床研究所
△通信作者:侯艳,E-mail:houyan@ bjmu.edu.cn

方法的历史可以追溯到 2002 年，Ideker 等[17]整合了酵母蛋白质-蛋白质和蛋白质-DNA 的相互作用，使用模拟退火搜索算法，对子网络进行评分排序，识别了重要的调控子网络，揭示了重要的信号传导与调控通路，拉开了模块识别方法发展的序幕。现在已经有许多模块识别方法被开发出来，其主要原理可以分为三类：基于矩阵分解的方法、基于偏最小二乘的方法和基于网络的方法。

1.基于矩阵分解的方法

基于矩阵分解的模块识别方法将组学矩阵分解为系数矩阵与模块矩阵，后两者具有实际意义，模块矩阵表示模块包含的基因，而系数矩阵表示不同模块在各患者上的表达情况（图 1），因此一般是非负的，但解不一定唯一。

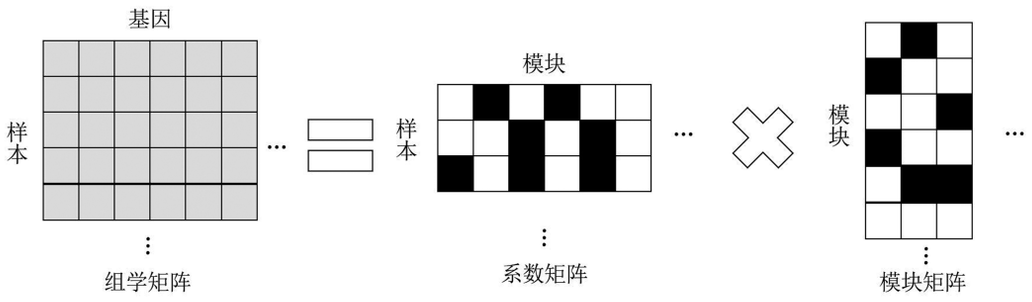

图 1　基于矩阵分解的模块识别方法

为了得到稳定的生物学解释，Lee 等[18]提出了 NMF（non-negative matrix factorization）方法，其表示如下：

$$min_{W,H} \| X-WH \|_F^2$$
$$s.t. W \geqslant 0, H \geqslant 0. \quad (1)$$

NMF 最小化分解前后原组学矩阵 X 与系数矩阵 W 模块矩阵 H 乘积之差，该差由 Frobenius 范数衡量，由此得到确定的系数矩阵与模块矩阵，其数值分别代表模块中各基因的表达情况和患者体内各模块的激活的情况，直观描述了生物学上的关系。Zhang 等[19]在此基础上将矩阵分解扩展至多组学，提出了 jNMF（joint non-negative matrix factorization）方法，其表示如下：

$$min_{W,H_1,\cdots,H_K} \sum_{k=1}^{K} \| X_k - W H_k \|_F^2$$
$$s.t. W \geqslant 0, H_k \geqslant 0, k=1,\cdots,K. \quad (2)$$

在 K 个组学中分别分解各个组学矩阵X_k并最小化分解前后之差，其采用一个共同的系数矩阵 W 和每个组学各自的模块矩阵H_k将各组学联系起来，每个组学各自的模块矩阵表示模块在该组学中包含的组学分子，而共同的系数矩阵表示多组学模块在各患者上表达情况。在肿瘤与癌症基因组图谱（the cancer genome atlas，TCGA）的卵巢癌样本中，jNMF 通过整合甲基化、基因表达、miRNA 表达数据得到了多组学的生物学模块，揭示了一些在单组学分析中被忽视的调控通路和跨组学的分子联系，这些模块与不同的患者亚型相关联，具有较好的生物学解释性。然而多组学间具有异质性，不同组学具有不同的分布特点，对于某一特定疾病的重要性也不同，jNMF 没有考虑到这一点。Yang 等[20]对其进行了改进，提出了 iNMF（integrative non-negative matrix factorization）方法，其表示如下：

$$min_{W,H_1,\cdots,H_K,V_1,\cdots,V_K} \sum_{k=1}^{K} \| X_k - (W + V_k) H_k \|_F^2 + \lambda \sum_{k=1}^{K} \| V_k H_k \|_F^2$$
$$s.t. W \geqslant 0, H_k \geqslant 0, V_k \geqslant 0, k=1,\cdots,K. \quad (3)$$

虽然还是对 K 个组学分别进行分解，但将其中共同的系数矩阵分为共同部分 W 和各组学独立部分V_k，并对独立部分V_kH_k额外施加 Frobenius 范数惩罚项以保证共同部分保留了足够的信息量，λ 为独立部分的可调节惩罚系数。该方法考虑了多组学整合时组学间的共同点与异质性，在卵巢癌的应用中，iNMF 通过整合 DNA 甲基化、基因表达和 miRNA 表达数据发现了癌症相关的通路和与亚型相关的模块，有助于癌症发生发展机制的理解。

虽然 NMF 产生较早，但近年来也有不少模块识别方法是基于矩阵分解开发的，如 Moon 等[21]提出的 JDSNMF（joint deep semi-non-negative matrix factorization）结合深度学习技术对多组学数据进行联合多层分解，将深度神经网络引入组学分析，多层非线性的激活函数对组学数据具有强大的拟合能力，能有效使用大数据进行高效计算，同时也利用了矩阵分解思想的生物学解释性，在阿尔茨海默症队列中识别了与年龄及亚型相关的模块，提供了很好的疾病机制研究线索。

2. 基于偏最小二乘的方法

基于偏最小二乘原理的方法为每个组学矩阵考虑一个载荷系数，将组学矩阵转化到潜变量空间，最大化潜变量间的协方差得到载荷系数，协方差大表明数据结构分布相似，表征了组学间的共性结构，此时各组学的载荷系数代表了模块构成的情况（图 2）。

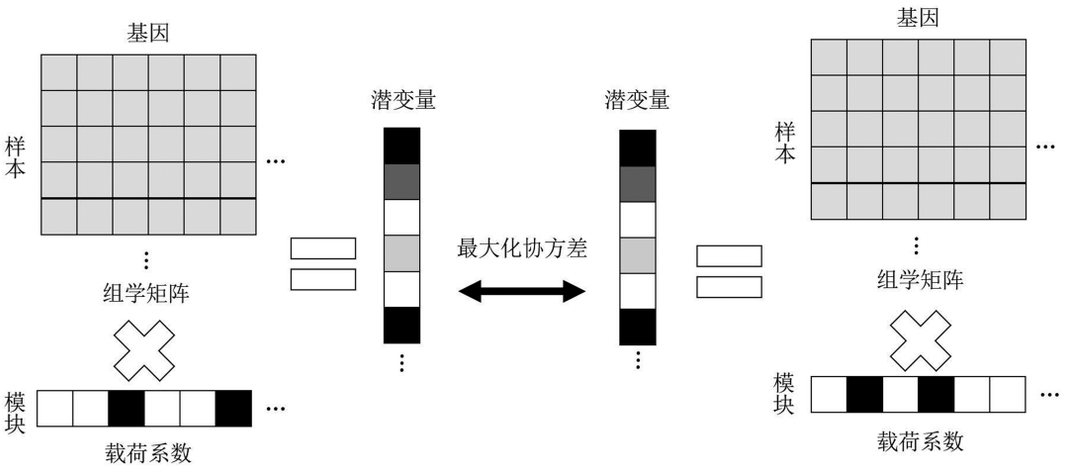

图 2 基于偏最小二乘的模块识别方法

传统的偏最小二乘方法表示如下，其中 X 为组学矩阵，T 为潜变量，P 为载荷系数：

$$T_1=X_1P_1;T_2=X_2P_2$$

$$\max cov(T_1,T_2),s.t.\ \|P_1\|=1,\|P_2\|=1 \quad (4)$$

偏最小二乘方法一次只能提取一个生物学模块，需要从数据矩阵中减去前一组载荷系数和潜变量信息，再继续进行偏最小二乘才能得到剩下的模块。最大化协方差使该类方法一般只能应用于两个组学，为了同时在更多的组学种类中应用，Li 等[22]提出 sMBPLS（sparse multi-block partial least squares）方法，把偏最小二乘扩展至多个组学，其表示如下：

$$\max cov(T,U)-2\lambda\sum_{k=1}^{3}|P_k|-2\lambda|Q|$$

$$T_k=X_kP_k,T=\sum_{k=1}^{3}B_kT_k,U=X_4Q$$

$$s.t.\ \|P_k\|=1,\|B_k\|=1,\|Q\|=1,k=1,2,3. \quad (5)$$

将 k 个组学分为两类，最重要的基因表达组学 X_4 为一类，其他组学 X_k 为一类，将其他组学的潜变量 T_k 乘上权重 B_k 加合为一个综合潜变量 T，通过使基因表达组学的潜变量 U 与综合潜变量 T 之间的协方差最大得到载荷系数与权重，权重表明了该组学在综合潜变量中的贡献，有助于识别对于特定疾病或患者最重要的组学成分，使结果更具有鲁棒性。此外该方法还在载荷系数 P_k、Q 上施加了 lasso 惩罚项以得到稀疏解，λ 为可调节的惩罚系数，使载荷系数 P_k、Q 是否非零能够反映模块包含基因的情况。sMBPLS 在 TCGA 卵巢癌的应用中，整合了拷贝数、DNA 甲基化、基因表达和 miRNA 表达数据识别了跨组学的显著功能富集模块，基于这些发现可以研究癌症相关机制，重建跨组学分子调控网络。近年来，Vahabi 等[23]在 sMBPLS 的基础上继续进行改进，提出 Cox-sMBPLS 方法，整合了表观基因组学、基因组学和转录组学，并应用于有监督的生存预测上，选择前几组潜变量作为生物学模块，在心力衰竭患者队列中发现了影响生存的模块并探索了生物学机制，可以用于新的生物标志物发现。

3. 基于网络的方法

相较于前两类方法，基于网络的模块识别方法更加直观。在网络方法中，点代表组学分子，边代表组学分子间的相互作用，赋予了网络实际生物学意义，可以直接表示组学分子间的调控关系，功能模块以紧密连接的子网络形式存在。因此网络方法一般分为两步，构建网络与识别模块。构建网络有两种方法，一是使用现成的生物学网络，如蛋白质互作网络；二是基于数据构建网络，网络的边由组学分子间联系紧密性的度量表示，如相关系数。识别模块是对于网络图形性质的分析，目的是找出网络中联通程度最大的区域作为模块，具体方法比如随机游走，或定义一个代表模块内相互作用强度的评分，随机搜索评分最高的模块构成方式，得到重要的子网络，也就是调控模块（图 3）。

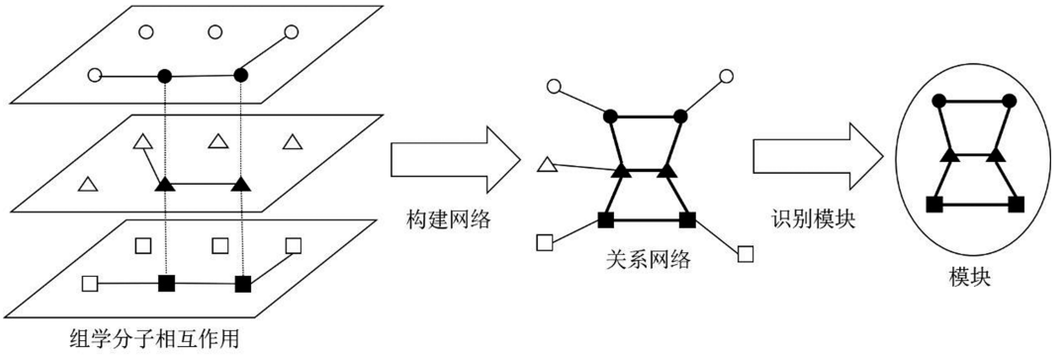

图 3 基于网络的模块识别方法

目前最经典的网络方法为加权基因共表达网络分析（weighted correlation network analysis, WGCNA）[24]，其使用 Pearson 相关系数来计算所有样本或条件下基因之间的相关性，再通过层次聚类识别网络中的共表达模块，将模块作为整体分析其与表型的关联。基因共表达模块能较好的解释癌症机制，在表型预测、生物标志物或治疗靶点发现中具有重要作用。如 2021 年 Long 等[25]使用该方法在胆管癌患者中根据基因表达数据识别了 7 个与临床特征显著相关的模块，模块中的部分基因是患者预后的重要生物标志物，提示了与肿瘤分化相关的机制。WGCNA 提出至今也经历了不少发展，主要体现在使用场景扩展和方法改进上。对于使用场景，WGCNA 最初仅应用于基因表达，现在已经扩展到蛋白质[26]、代谢[27]等组学上。方法上的改进如 Tesson 等[28]认为简单的 Pearson 相关不足以表示两个基因间相互作用的相似程度，而改用拓扑重叠分数，其考虑了与这两个基因有相互作用的所有基因间重叠程度来描述共表达。该研究提出 DiffCoEx 方法，使用 Pearson 相关或 Spearman 相关计算邻接矩阵，邻接矩阵表示节点间的相邻关系，由不同条件下的邻接矩阵得到邻接差异矩阵，高的邻接差异表明基因间的共表达状态在两种条件下差异较大，再从邻接差异矩阵的拓扑重叠分数中推导出差异网络，然后使用层次聚类得到差异共表达模块。2021 年 Ghobadi 等[29]使用 DiffCoEx 在淋巴瘤患者、HTLV-1 病毒携带者与正常人群中识别了不同的共表达模块，发现了特定的疾病基因通路在不同人群中活性的差异，其中一些基因能够作为可能的生物标志物或治疗靶点。除了 WGCNA 一类方法，还有其他的网络方法如 Zhang 等[30]提出的 iMCMC，利用特征间的相互作用的相似程度，分别定义了变异网络和表达网络，标准化后合并为综合网络，最后识别模块时要求边的权重和最大的同时包含的节点尽量少。iMCMC 在胶质母细胞瘤和卵巢癌的应用中发现的核心模块包含一些已知的通路并富集了致癌抑癌基因，在癌症机制研究中发挥重要作用。

网络结构本身具有生物学意义使得已知生物学信息的加入非常方便，这成为网络方法的一大优点。利用数据库中的调控关系如通路信息、蛋白质互作关系作为网络框架的参考，可以更新或补充仅基于样本数据构建的网络结构。如 Webber 等[5]提出的 MAGNETIC（modular analysis of genomic NETworks in cancer）方法，使用标准化的拷贝数、甲基化、体细胞突变、基因表达和蛋白质丰度数据作为输入，计算组学间与组学内所有特征间的相关性作为网络的边，为了让不同种类组学分子间的相关性可比，以数据库里的蛋白质互作关系为基准进行了标准化，构造了多组学的分子相似性网络。接下来使用随机游走算法对网络进行图聚类，识别了包含多组学分子的模块，并根据模块表达活性评分，其分数可以作为生物标志物，反映重要的细胞特征，如免疫细胞浸润情况等，并通过在细胞系中得到的药物反应数据分析模块的药物相互作用，使模块分数成为稳健的药物反应预测因子。

4.其他方法

除了以上三类主流方法，还有不少基于其他原理的方法，如 Park 等[31]使用稀疏重叠的组 lasso 识别癌症中的关键模块，该方法同时在模块间与模块内施加稀疏限制，不仅减少了过拟合，而且也考虑到了一个基因出现在多个模块中的情况。该方法使用已知的通路信息作为先验知识，在 TCGA 的五个癌症数据集中整合了基因表达、拷贝数变异、突变状态、蛋白质互作关系、致癌/抑癌基因信息，有效识别了重要的癌症模块，富集了癌症驱动基因与相互作用，为癌症机制的研究提供了方向。Silverbush 等[4]提出的 ModulOmics，使用 TCGA 乳腺癌数据集，整合蛋白质互作关系、突变状态、转录共调控和 RNA 共表达数据，为每个模块定义了基于分子间相互作用的评分。为了更好的搜索并评分模块，ModulOmics 采用了两步优化过程来节省计算量，先使用整数线性规划初步识别可能的跨组学模块，再通过随机搜索使模块评分最高，确定最优模块，其在乳腺癌数据集上发现的模块富集了功能相关的癌症驱动基因，与乳腺癌的不同亚型相关联，揭示了乳腺癌各亚型发展的驱动机制。Rappoport 等[32]使用聚类的思想识别模块，该研究指出以前的聚类方法都建立在各组学有一个共同底层结构的假设上，而这个假设不一定在所有情况下都成立，由此提出了 MONET 根据患者间相似性进行聚类，通过图融合寻找基于患者相似性构建的各组学网络的重叠子图来识别模块，其得到的模块可能只包含部分组学，且可能使某些患者不能识别出任何模块，但其区分患者亚型的能力优于其他聚类方法，发现的模块也具有很强的生物学和临床相关性，还能够让 MONET 适应存在组学缺失的数据。该方法的多项应用中，在卵巢癌中识别的模块可用于研究生物学机制且与不同的患者亚型高度相关；对 TCGA 的 10 种癌症进行的亚型聚类，在调整兰德指数（adjusted rand index，ARI）等聚类指标、癌症分期等亚型临床标签富集水平等结果上都获得了较好的效果；在单细胞组学和发育组学的数据集上发现的模块很好地描述了不同类型细胞的功能，能够区分细胞类型并解释细胞分化的机制。

在三类主要的模块识别方法中，基于矩阵分解的方法直观地将组学矩阵分解为模块矩阵与系数矩阵以识别模块，其与深度学习的结合能够更好地拟合非线性关系。基于偏最小二乘的方法通过最大化组学潜变

量间协方差,计算各组学的载荷系数得到模块,但一般只在两个组学中实现,在更多的组学中实施该方法时需对其他组学进行处理。基于网络的方法包括构建网络和识别模块两步,对生物学先验知识有较好的适应性。前两类方法需要主观设定模块个数,而网络方法是根据合适的阈值识别子网络,更加客观。而基于其他原理的模块识别方法扩展了模块识别的思路,共同丰富了模块识别领域的研究,成为模块识别进一步发展的基础。

模块识别方法评估

一般来说,评估模块识别方法的性能即评估是否识别了正确的模块构成,分为在模拟数据中和真实数据中两种情况。使用模拟数据评估的优势为可以事先设定相互作用的分子组成模块,与发现的模块相比较,使用准确度、召回率或超几何检验的 p 值等指标能够较好地反映方法识别模块的能力,但在模拟数据中设置模块信息具有难度而且模拟数据是数据的理想情况,与真实数据存在差别。使用真实数据评估能够反映方法实际应用的情况,其中最常用的评估方法是将识别的模块与生物过程联系起来[33],具体做法是使用 KEGG 或 GO 等数据库中现成的模块进行功能富集,在为模块提供功能注释的同时也能够通过超几何检验的 p 值、f-measure 或目标检测中的平均精度等指标比较数据库中的模块和识别的模块。如 Levi 等[1]对 6 种基于网络的单组学模块识别方法进行了评估比较,使用 GO 进行功能富集,通过超几何检验的 p 值衡量方法的效果,该研究指出基于网络的方法较多受先验网络信息影响,忽视了数据的特定生物学背景,提出在大量的置换数据上运行识别方法获得 p 值的背景分布,然后从真实数据得到的 p 值中去掉背景分布的方法来校正偏倚,实现了不同网络模块识别方法的比较。但该研究只评估了基于网络的模块识别方法,其他原理的方法评估是否适用还需要进一步研究。然而使用真实数据评估也存在缺点,由于患者或疾病的异质性,在特定的数据中很难找到其模块构成的金标准。此外,数据库里的通路信息覆盖基因较少,难以评估不在通路中的基因构成的模块。对于这些新发现的模块,只能通过实践进行检验,如利用其进行表型的预测或亚型的分类,利用相应的准确性指标证明模块在不同患者中具有实际的区分效果。当然,要确切地验证模块中的相互作用或模块影响表型的机制,进行相关的生物学实验是最终且证据力度最高的选择。

未来展望

模块思想的出现是组学分析发展的重要一步,特别是在生物学机制研究方面,但该领域的发展也面临不少的挑战。目前,生物学模块识别方法数量还相对较少,不同方法结果间一致性不高,表明其关注组学数据不同方面的特点。同时很多方法仅由数据驱动,未使用生物学信息,考虑到目前在生物学上的了解有限,结合生物学信息和数据本身特点,探索并更新生物学知识的方法更加可取。细胞的生理活动是随时变化的,模块的活性也处于动态变化中,需要多时间点的数据和利用时序的方法在动态过程中提取模块,观察其中的变化过程与因果关系,探究动态的疾病机制,这对于癌症的进展与药物治疗反应的研究都十分有用。最近 Bodein 等[34]使用多时间点的多组学数据构建了表达谱网络,通过随机游走识别子网络模块,识别了关键功能模块和动力学作用机制,为动态模块识别做出探索。此外,模块识别方法的评估还没有形成系统的标准且涉及的方法在原理上也不够全面,在缺少特定数据集金标准的情形下,开发适用于不同原理方法的基准,形成统一的评估体系,以选择不同条件下最合适的识别方法是未来的一大挑战。

组学分子相互作用呈现无标度网络分布,其中关键节点对外界扰动有强大承受能力,但面对协同干扰则会显得脆弱,所以在模块中多个点同时展开干预以治疗疾病效果较好,凸显了正确识别模块的重要性。此外,组学分析中利用生物学模块的模型往往兼具好的效果与生物学解释性,如一些深度学习模型使用组学分子所属模块的信息构建稀疏连接,使神经网络节点具有实际意义,改善了深度学习模型的解释性,同时也减少了过拟合,提高了泛化能力[35-36]。虽然基于模块的组学分析方法还在发展中,但其出现表明研究者开始以整体的角度观察一组分子的变化对表型的影响,从根源上减少了假阳性与假阴性,其描述生理功能与微环境的方式也便于生物学机制的探索,由此可以更有效地进行表型或生存预测、疾病亚型区分与生物标志物发现等一系列临床转化。

参 考 文 献

[1] Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls[J]. Molecular Systems Biology, 2021, 17(1): e9593.

[2] Li F, Wu T, Xu Y, et al. A comprehensive overview of oncogenic pathways in human cancer[J]. Briefings in Bioinformatics, 2020, 21(3): 957-969.

[3] Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis[J]. Computational and Structural Biotechnology Journal, 2021, 19: 949-960.

[4] Silverbush D, Cristea S, Yanovich-Arad G, et al. Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules[J]. Cell Systems, 2019, 8(5): 456-466.

[5] Webber JT, Kaushik S, Bandyopadhyay S. Integration of Tumor

Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics[J]. Cell Systems, 2018, 7(5): 526-536.

[6] Gustafsson M, Nestor CE, Zhang H, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis[J]. Genome Medicine, 2014, 6(10): 82.

[7] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning[J]. Briefings in Bioinformatics, 2022, 23(1): bbab454.

[8] Staiger C, Cadot S, Györffy B, et al. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis[J]. Frontiers in Genetics, 2013, 4: 289.

[9] Azad AK, Lee H. Voting-based cancer module identification by combining topological and data-driven properties[J]. PLoS One, 2013, 8(8): e70498.

[10] Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations[J]. Nature Genetics, 2018, 50(9): 1219-1224.

[11] Qiu YQ, Zhang S, Zhang XS, et al. Detecting disease associated modules and prioritizing active genes based on high throughput data [J]. BMC Bioinformatics, 2010, 11: 26.

[12] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(43): 15545-15550.

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

[14] Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Research, 2017, 45(D1): D353-D361.

[15] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong[J]. Nucleic Acids Research, 2019, 47(D1): D330-D338.

[16] Van Kampen AH, Moerland PD. Taking Bioinformatics to Systems Medicine[J]. Methods in Molecular Biology, 2016, 1386: 17-41.

[17] Ideker T, Ozier O, Schwikowski B, et al. Discovering regulatory and signalling circuits in molecular interaction networks[J]. Bioinformatics, 2002, 18 Suppl 1: S233-S240.

[18] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.

[19] Zhang S, Liu CC, Li W, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data[J]. Nucleic Acids Research, 2012, 40(19): 9379-9391.

[20] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data[J]. Bioinformatics, 2016, 32(1): 1-8.

[21] Moon S, Lee H. JDSNMF: Joint Deep Semi-Non-Negative Matrix Factorization for Learning Integrative Representation of Molecular Signals in Alzheimer's Disease[J]. Journal of Personalized Medicine, 2021, 11(8): 686.

[22] Li W, Zhang S, Liu CC, et al. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data[J]. Bioinformatics, 2012, 28(19): 2458-2466.

[23] Vahabi N, Mcdonough CW, Desai AA, et al. Cox-sMBPLS: An Algorithm for Disease Survival Prediction and Multi-Omics Module Discovery Incorporating Cis-Regulatory Quantitative Effects[J]. Frontiers in Genetics, 2021, 12: 701405.

[24] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9: 559.

[25] Long J, Huang S, Bai Y, et al. Transcriptional landscape of cholangiocarcinoma revealed by weighted gene coexpression network analysis[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa224.

[26] Gutierrez-Quiceno L, Dammer EB, Johnson AG, et al. A proteomic network approach resolves stage-specific molecular phenotypes in chronic traumatic encephalopathy[J]. Molecular Neurodegeneration, 2021, 16(1): 40.

[27] Bhargava P, Fitzgerald KC, Calabresi PA, et al. Metabolic alterations in multiple sclerosis and the impact of vitamin D supplementation[J]. JCI Insight, 2017, 2(19): e95302.

[28] Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules[J]. BMC Bioinformatics, 2010, 11: 497.

[29] Zarei Ghobadi M, Emamzadeh R, Teymoori-Rad M, et al. Decoding pathogenesis factors involved in the progression of ATLL or HAM/TSP after infection by HTLV-1 through a systems virology study[J]. Virology Journal, 2021, 18(1): 175.

[30] Zhang J, Zhang S, Wang Y, et al. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data[J]. BMC Systems Biology, 2013, 7 Suppl 2: S4.

[31] Park H, Niida A, Miyano S, et al. Sparse overlapping group lasso for integrative multi-omics analysis[J]. Journal of Computational Biology: a Journal of Computational Molecular Cell Biology, 2015, 22(2): 73-84.

[32] Rappoport N, Safra R, Shamir R. MONET: Multi-omic module discovery by omic selection[J]. PLoS Computational Biology, 2020, 16(9): e1008182.

[33] Barel G, Herwig R. Network and Pathway Analysis of Toxicogenomics Data[J]. Frontiers in Genetics, 2018, 9: 484.

[34] Bodein A, Scott-Boyer MP, Perin O, et al. Interpretation of network-based integration from multi-omics longitudinal data[J]. Nucleic Acids Research, 2022, 50(5): e27.

[35] Zhao L, Dong Q, Luo C, et al. DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis[J]. Computational and Structural Biotechnology Journal, 2021, 19: 2719-2725.

[36] Seninge L, Anastopoulos I, Ding H, et al. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics[J]. Nature Communications, 2021, 12(1): 5684.

（责任编辑：邓　妍）