中国科学 信息科学
SCIENTIA SINICA Informationis

基于生物网络的关系推断原理、方法与应用

李梢[1], 李衍达[2], 古槿[3], 张鹏[4] and 江瑞[4]

Citation: 中国科学: 信息科学 (2021); doi: 10.1360/SSI-2021-0243

View online: https://engine.scichina.com/doi/10.1360/SSI-2021-0243

Published by the 《中国科学》杂志社

中国科学: 信息科学

SCIENTIA SINICA Informationis

论文

基于生物网络的关系推断原理、方法与应用

李梢[1*], 张鹏[1], 古槿[1], 江瑞[1], 李衍达[1*]

1. 清华大学自动化系, 清华大学北京市中医药交叉研究所, 北京信息国家研究中心, 北京 100084

* 通信作者. E-mail: shaoli@tsinghua.edu.cn, daulyd@tsinghua.edu.cn

国家自然科学基金项目 (批准号: 62061160369, 81630103, 81225025) 资助项目

摘要 在生物医学大数据时代, 如何全面有效地发现致病基因、药物靶标等关键要素, 从整体上理解宏观表型的微观本质, 是目前信息科学与中西医学交叉研究面临的重大共性挑战之一。生物系统是典型的复杂系统, 克服上述挑战的关键是: 如何通过深入理解复杂生物系统的“关系”本质, 解决复杂系统多层次信息融合难题以及生物医学大数据中广泛存在的维度高、噪声大、样本少等难点。“生物网络”是构成复杂生物系统的基础, 反映人体内部基因和基因产物等各种生物分子的相互关系、生物分子与疾病和药物等不同层次的关系, 生物网络已被广泛用于生物医学大数据的分析。李梢课题组从 20 余年前开始中西医药与生物网络的关联研究, 率先提出“网络靶标”假说, 并进行方法构建与应用。本文对基于生物网络的关系推断理论与方法做一总结与思考。首先, 从原理上, 发现疾病宏观表型与微观分子在复杂生物网络中存在“层次模块化关系”, 即宏观层次的涌现在微观上具有局域模块性, 宏观表型越相似, 微观致病基因或药物靶标在网络上的模块性关联越强。其次, 从方法上, 给出基于生物网络从生物医学大数据、少量目标样本中推断关键生物要素的“关系推断”一般性方法框架: 以层次模块化关系为基础, 从全局角度进行关系网络构建、关系表示与建模、未知关系推断, 实现关系的实体化、数学化、整体化。进而, 从应用上, 基于生物网络的关系推断方法在致病基因与药物靶标预测、疾病标志物识别、中医药机制解析等方面表现出很好的性能。综上, 关系推断方法能够为从系统角度和分子水平揭示中西医药科学原理提供系统解决方案, 也为网络药理学等新兴学科提供重要的原理和方法学支撑。

关键词 复杂生物网络, 模块化, 多层次关联, 关系推断, 小样本推断, 网络药理学

1 基于生物网络的关系推断的提出背景

随着高通量组学数据的积累与信息处理技术的发展, 医学生命科学的研究思维正发生从还原论向系统论的变革, 其中一个重要标志便是从复杂系统和生物网络的角度来理解疾病诊疗的内在机理 [1,2]. 无论是疾病发生发展还是中西药物对其干预调节, 都与表征人体复杂系统的大规模生物网络密切相关. ” 复杂生物网络” 也是利用信息科学理解人体复杂系统运行规律, 实现中西医药与信息科学交叉创新的一个重要突破点. 而复杂生物网络研究的一个根本问题便是对于生物要素之间各种复杂” 关系” 的理解, 包括” 关系” 的实体化, 数学化, 整体化, 这也是从根本上解析” 整体大于局部之和” 这一系统论

基本规律的核心. 可以认为, 中医与西医, 还原论与系统论的一个本质区别, 即是对于” 关系” 的理解. 中医药的特点在于整体观, 即从” 整体” 的视角看待病证表型与药物之间的复杂” 关系”, 但缺少与分子, 细胞等微观元素之间关系的解析; 现代医学的兴起则与还原论研究模式息息相关, 但是缺乏对于生命整体, 系统的把握. 如何系统地衔接宏观整体与微观实体, 成为目前中西医学共同面临的一个核心问题, 也是信息科学的一个前沿问题. Science 创刊 125 周年之际, 也将”How will big pictures emerge from a sea of biological data?” 列为 25 个人类未知的大科学问题之一, 体现了理解复杂生物数据之间的” 关系” 以及整体涌现机制的重要性. 因此, 从系统角度理解疾病中西医诊疗内在规律的一个重大挑战, 便是如何解析大规模复杂生物网络中的各种” 关系”, 进而推断致病基因, 药物靶标等关键要素.

针对上述重大挑战, 常规基于模型驱动的机器学习或基于数据驱动的统计推断范式在解析复杂生物系统中的” 关系” 时往往会面临以下瓶颈问题: 1) 多层次信息难以融合. 在复杂生物网络分析中, ” 多层次” 既包括疾病表型, 中西药物等宏观信息, 也包括致病基因, 药物靶点等生物分子, 细胞等微观信息. 随着多种测序技术的深入发展, 对复杂生物系统微观信息的探测又可进一步分为基因组, 转录组, 蛋白质组, 表观组等多个微观层次. 宏微观多层次信息的有效融合是理解复杂生物系统” 关系” 的基础与前提. 另一方面, 复杂疾病相关的生物网络往往具有较大的规模, 属于大尺度生物网络, 其高度的复杂性决定了那些基于小尺度网络所建立的信息融合模型往往难以直接应用. 因此, 如何建立针对大尺度, 多层次生物网络的信息融合模型, 是关键要素推断面临的一个难题. 2) ” 小样本推断” 难题. 在生物医学数据中, 小样本问题广泛存在. 以疾病基因关系推断为例, 目前已知致病基因的表型 (即正样本) 所占比例非常小, 存在大量无已知基因的疾病表型 (例如中医表型) . 即使是疾病 - 基因关系已知, 也存在负样本难以界定的情况. 正样本少乃至缺失, 负样本不明确, 对目前的机器学习, 人工智能分析模式形成很大挑战. 如何突破样本标签的局限, 实现基于小样本甚至零样本的” 学习” 是关键要素推断时面临的另一个难题. 因此, 复杂生物网络” 关系” 解析的一般性理论与方法尚未形成, 迫切需要理论与方法的突破.

人擅长于知识抽象与类比推理, 在遇到” 复杂” 问题时, 即使只有少量” 样本”, 仍可调动与问题相关的各种信息, 案例, 知识帮助问题的解决与决策. 中医学在遇到新冠肺炎等新发疾病时也能进行诊治, 其原理也在于” 取类比象” 的整体思维. 人的抽象与类比思维, 以往多用于宏观层面, 如果有效用于微观层次, 则能为我们克服多层次信息融合难题和小样本推断难题, 实现关键要素的精准推断提供全新思路. 在这个方面, 国内外学者开展了有益的尝试, 取得了一些重要突破. 我国学者李梢 1999 年率先提出中医药与分子网络相关的科学假说 [3], 并尝试将中医” 取类比象” 思维用于复杂生物网络分析, 创建基于宏, 微观类比规律的” 关系推断” 理论与方法, 为解决中西医药分子机制等典型小样本推断难题提供了突破口 [4,5]. 2007 年英国学者 Hopkins 提出” 网络药理学”, 强调了基于生物分子网络的药物靶点类比分析对于新靶点推断和药物作用机理解析的重要性, 并认为是” 下一代药物发现模式”. 2017 年, 美国科学院院士 Loscalzo 在 ¡新英格兰医学¿ 发表社论 [1], 回顾了从 18 世纪以来还原论在医学领域的发展历程与局限性, 再次强调了以多层次生物网络为基础的” 网络医学” 对于理解人体复杂系统以及医学发展的重大意义. 这些研究为从生物网络角度将抽象与类比思维应用于微观层次, 进而建立全新的关系推断方法带来了曙光.

2 基于生物网络的关系推断的原理与方法

从大规模复杂生物网络推断微观上的致病基因, 药物靶标等关键要素, 其前提是理解要素之间的复杂” 关系”. 基于对” 关系” 的理解, 李梢课题组提出不同于当前” 单基因, 单靶标” 还原论模式的”

生物网络, 网络靶标" 研究模式 [3~6], 强调了对于" 关系" 规律发现, 定性定量描述, 及其在研究复杂疾病机制, 揭示中医药科学基础上的重要性. 通过对疾病诊疗机制的观察以及相关数据的全面分析, 课题组发现疾病表型 - 基因以及药物 - 靶标在复杂生物网络中广泛存在" 层次模块化关系", 表现为" 模块化构成" 和" 多层次关联" 这两个方面.

以致病基因预测为例, 首先, 复杂疾病相关的致病基因数量是有限的 (截止 2021 年 5 月, COSMIC 数据库中记录的肿瘤致癌突变基因总数仅为 576 个) , 针对少数节点的干预就能改变疾病表型. 即使是具有复杂成分的中药, 其有效成分及其作用的潜在干预靶标也是有限的. 更为重要的是, 这些致病基因之间, 干预靶标之间存在高度的相似性, 在生物网络中倾向于形成紧密相连的" 网络模块", 呈现局域聚集的分布特征. 这提示, 复杂系统宏观层次的涌现在微观上具有" 模块化构成" 的特点. 这里的模块化主要是指, 模块内部紧密关联, 而模块与周围要素之间的关联相对较弱; 进一步地, 我们还观察到宏, 微观不同层次模块之间也存在关联, 表现为宏观表型越相似, 相应致病基因或药物靶标在网络上的模块性关联越强, 提示不同层次生物要素之间存在基于模块性构成的" 多层次关联". 生物要素在网络中形成的这种由模块化构成和多层次模块关联构成的关系, 我们将其归纳为" 层次模块化关系".

层次模块化关系为解决复杂生物网络关键要素推断所面临的多层次信息融合难题, " 小样本推断" 难题提供了基础. 一方面, 我们发现疾病复杂生物网络具有模块性构成规律, 这提示在对大尺度网络分析时, 可以聚焦局域模块, 大幅降低分析的复杂度. 更为重要的是, 利用网络模块, 可以实现复杂生物系统宏观 (如中西医表型) 与微观 (如生物分子) 信息的统一表示, 而模块之间的多层次关联规律 (如相似性) 又可以进一步用于宏, 微观信息的建模分析, 从而实现多层次信息融合; 另一方面, 利用要素在局域模块内与其他要素的相似性, 以及模块之间的关联关系, 我们可以从关注要素自身的特征转到关注要素之间的" 关系" 上, 将与其他要素的关系作为该要素新维度的特征. 这样, 对于任意一个要素 (或样本) , 无论其是否具有先验的标签信息, 都可以建立其与正样本之间的相似关系, 实现计算推断, 从而使得小样本甚至零样本推断成为可能. 因此, 以层次模块化关系为基础, 有望建立复杂生物网络要素推断的新方法, 突破复杂生物网络解析的方法学瓶颈.

需要指出的是, 网络" 模块" 性也是复杂网络与复杂系统可控性分析的关键点. 例如, Liu 等人揭示了任意复杂有向网络的可控性与部分驱动节点集密切相关 [7]; Vinayagam 等人对发现了控制大规模生物相关网络 (即人类 PPI 网络) 所需的最少驱动节点, 确定了部分节点在调节健康和疾病状态之间的过渡中起关键作用 [8]. Yan 等通过将线虫神经系统模拟成一个有向网络, 构建动力学方程对每个线虫神经元在运动行为中的参与进行预测 [9]. 复杂系统可控性分析通过探索复杂网络中结构 - 功能关系, 为从" 模块性" 角度理解复杂系统要素之间的" 关系" 提供有力的理论支撑.

于是, 我们总结出一种通过网络将" 关系" 实体化, 数学化, 整体化, 进而推断系统关键要素的策略, 称之为" 关系推断", 分为三个主要步骤: 1) 关系网络构建: 即从文献, 生物实验或组学大数据中获取复杂系统要素之间的关系, 包括宏观层次要素之间的关系, 微观层次要素之间的关系以及宏 - 微观要素之间的关系. 生物网络便是复杂生物系统要素之间关系的常用表示方式; 2) 关系表示与建模: 对复杂系统宏 - 微观要素关系进行数学建模, 包含模块局域的界定以及模块之间关联关系的函数表示. 值得注意的是, 所建立的函数需表示整个复杂系统宏, 微观要素模块之间的关系, 即需体现推断的全局性. 这是关系推断方法能反映复杂系统整体性的核心. 3) 未知关系推断: 即利用 2) 中所建立的函数表示, 充分利用复杂系统已知的要素关系去推断待定的关键要素.

关系推断方法的一般形式可以表示如下:

(i) 关系网络构建

关系推断的前提是如何将异构, 多源大数据转换为关系网络. 本文将这样一个关系网络定义为五

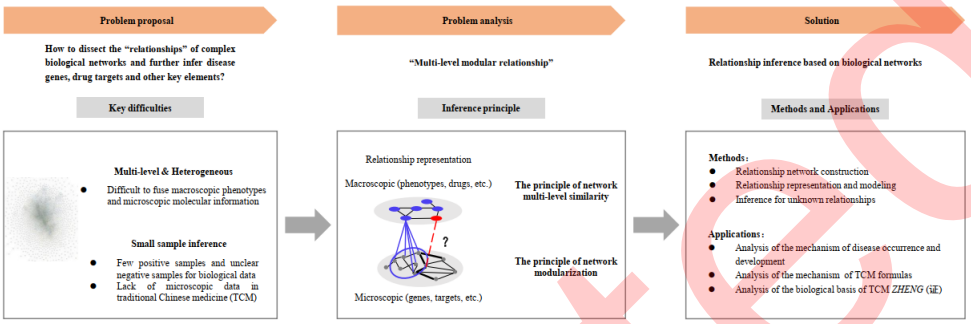

图 1　基于生物网络的关系推断问题提出, 分析与解决流程

Figure 1　The process of raising, analyzing and solving problems related to biological network relationship inference

元组 $G=(\boldsymbol{y},\boldsymbol{x},\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})$, 其中:

$\boldsymbol{y}=(y_i)_{n\times 1}$, 表示 n 个宏观层次要素 (如疾病表型等) 构成的向量;

$\boldsymbol{x}=(x_i)_{m\times 1}$, 表示 m 个微观层次要素 (如疾病基因, 干预靶标等) 构成的向量;

$\boldsymbol{U}=(u_{ij})_{n\times n}$, 表示 n 个宏观层次要素之间的关系,$u_{ij}$ 可以是二值关系 ($u_{ij}=0$ 或 1) 或数量关系 ($u_{ij}\in\mathbb{R}$) ;

$\boldsymbol{V}=(v_{ij})_{m\times m}$, 表示 m 个微观层次要素之间的关系, v_{ij} 可以是二值关系 ($v_{ij}=0$ 或 1) 或数量关系 ($v_{ij}\in\mathbb{R}$) ;

$\boldsymbol{W}=(w_{ij})_{n\times m}$, 表示宏观层次要素与微观层次要素之间的关系, w_{ij} 可以是二值关系 ($w_{ij}=0$ 或 1) 或数量关系 ($w_{ij}\in\mathbb{R}$) ;

根据以上定义, 可以得到:

二元组 $A=(\boldsymbol{y},\boldsymbol{U})$ 表示宏观要素之间的关系网络. 二元组 $B=(\boldsymbol{x},\boldsymbol{V})$ 表示微观要素之间的关系网络. 三元组 $C=(\boldsymbol{y},\boldsymbol{x},\boldsymbol{W})$ 表示宏观与微观要素之间的关系网络.

在以上定义中, 关系 G 既包含源于先验知识的已知关系 (如蛋白相互作用) , 也包含蕴藏在大数据中, 没有先验知识的" 新" 关系 (如共表达关联) . 此外, G 还可表示表型, 细胞, 分子等多个层次要素的关系.

关系网络构建的核心是网络中" 边" 的定义与度量. 生物要素之间的网络边关系包括共出现, 关联性, 因果性等多种类型, 可从功能, 形态, 表达等多个角度来进行度量. 其中, 共出现是指两个要素在文献或数据库中有直接或间接的关联关系记录; 关联性常通过将生物要素向量化表示与关联分析而得到的相似度量, 基因共表达关联便是一种常见的关联性度量; 而因果性既包括基因调控, 遗传相互作用以及药物对靶点干预等已知的因果关联, 也包括利用" 反事实推理" 等方法从生物大数据中获取的未知的因果关联关系, 例如 Park 等通过构建基于反事实推理的计算框架, 从单细胞转录组数据中推断出了具有因果性的疾病 - 基因关联 [10]. 此外, 网络边的度量也包含对其他" 隐藏" 在数据中的关系的获取, 其中常见的方式便是基于已知关系网络预测节点间关系.

(ii) 关系表示与建模

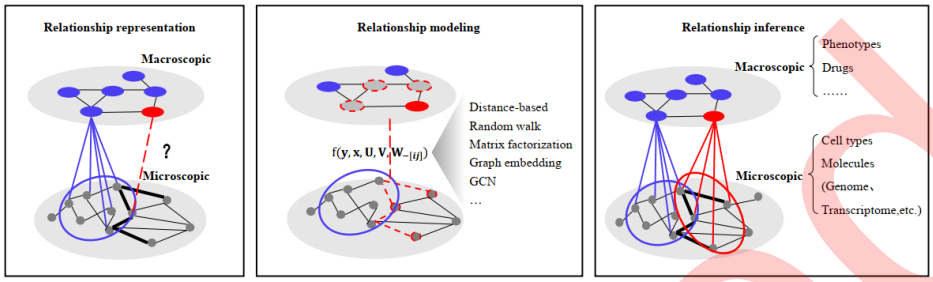

图 2　关系推断的一般性方法框架

Figure 2　The general methodological framework for relationship inference

利用复杂系统宏, 微观" 层次模块化关系", 关系建模可以分为两个部分:

a. 宏, 微观要素具有" 模块化构成" 规律:

$$\begin{aligned}\rho(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{C})&=\frac{\sum_{i=1}^{n_c}\sum_{j=1}^{m_c}1\left(\boldsymbol{W}_{ij}\neq 0\right)}{n_c\times m_c},\\ \rho(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{x}\in\boldsymbol{c}\text{ and }\boldsymbol{y}\in\boldsymbol{c})&\gg\rho(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{x}\notin\boldsymbol{c}\text{ or }\boldsymbol{y}\notin\boldsymbol{c}).\end{aligned}\tag{1}$$

在上面的公式中, ρ 表示任意微观要素和宏观要素共同构成的网络的稠密度. 其中 $\boldsymbol{C}$ 表示任意宏观与微观要素集合及其关系网络, $\boldsymbol{c}$ 表示具有模块化构成规律的宏微观要素集合及其关系网络. 网络模块化具有模块内部连接比较稠密, 模块与外部联系相对稀疏的特点.

b. 宏, 微观模块之间存在" 多层次关联":

$$w_{ij}=f\left(\boldsymbol{y},\boldsymbol{x},\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}_{-[ij]}\right).\tag{2}$$

其中, f 既可表示线性, 也可表示非线性. 上式通过 (i) 关系网络的构建, 对得到的关系网络进行建模, f 表示学习到的网络宏微观模块之间的关联. 这里侧重于通过已知关系, 对关系网络的整体规律进行建模和学习. 关系建模的一般性方法框架如图 2 所示.

多层次生物网络的关系建模, 主要有以下几种方法 (以疾病表型 - 致病基因关系模型构建为例, 下同) :

1) 基于网络距离的方法:

这类方法的假设是疾病表型 u 与致病基因 v 的关系强弱与相关节点模块之间的网络距离远近之间呈负相关关联 [11~13]. 通过构建网络节点之间的距离度量方式 d (如最短路径) , 即可定量建模不同层次模块之间的关联关系 W.

$$\begin{aligned}d(u,v)&=\text{ShortestPath}(u,v),\\ \boldsymbol{W}_{uv}&=e^{-d(u,v)}.\end{aligned}\tag{3}$$

2) 基于随机游走的方法:

这类方法将疾病表型, 致病基因等不同层次模块之间的关系建模为相关节点在网络中的转移概率 [14,15]. 随机游走方法可以应用于不同拓扑结构的网络, 为探索网络中的关系提供了一个有效的框

架.

$$\boldsymbol{p}^{t+1} = (1-\pi)\boldsymbol{W}^T\boldsymbol{p}^t + \pi\boldsymbol{p}^0. \tag{4}$$

在式 4 中, $\boldsymbol{p}$ 表示节点当前关联概率向量, $\boldsymbol{t}$ 为游走步数, 其中 $\boldsymbol{p}^0$ 为节点起始概率向量.

3) 基于矩阵分解的方法:

这类方法通常利用 PCA 或 NMF 等方法 [16~18] 将疾病表型 - 致病基因关系矩阵进行分解, 得到包含不同层次模块之间关系对的降维表示或向量表示 (如 meta-gene) . 如给定疾病表型的隐向量表示 $\boldsymbol{p}$ 和基因的隐向量 $\boldsymbol{q}$, 二者的关联分数 s 为:

$$s_{ij} = \boldsymbol{p}_i^T\boldsymbol{q}_j. \tag{5}$$

4) 基于图嵌入的方法:

这类方法通过将疾病表型 - 致病基因关联关系构建图模型, 并利用图嵌入方法将图模型映射到低维向量空间中, 实现不同层次模块关系的非线性建模. 图嵌入方法的核心目标是使得图中节点嵌入得到的向量形式应尽量保留图模型的结构信息和潜在的特性 [18~21], 这样保证了关系建模的可靠.

$$\begin{aligned}
&\max_f \sum_{u\in V} \log \Pr\left(N_s(u) \mid f(u)\right), \\
&\Pr\left(N_s(U) \mid f(u)\right) = \prod_{n_i\in N_s(u)} \Pr\left(n_i \mid f(u)\right), \\
&\Pr\left(n_i \mid f(u)\right) = \frac{\exp\left(f\left(n_i\right)\cdot f(u)\right)}{\sum_{v\in V}\exp(f(v)\cdot f(u))}.
\end{aligned} \tag{6}$$

在式 6 中, N_s 表示在 s 邻居采样策略下, 得到点 u 的邻居网络; $f(u)$ 表示节点特征函数; max 表示最大化节点 u 的特征表示, 观测到 $N_s(u)$ 的概率.

5) 基于图神经网络的方法

这类方法通过利用神经网络结构模型将疾病表型 - 致病基因的多层次关联关系转化为基于网络节点的隐向量之间的关系. 这类方法通常包含信息传递阶段和信息读出阶段. 信息传递阶段通常通过构建一个神经网络结构的模型, 获取表示关系节点的隐向量; 而读出阶段则依据上述隐向量来推断节点之间的关系 [5, 18].

$$\begin{aligned}
&\boldsymbol{m}_v^{t+1} = \sum_{u\in N_v} M_t\left(\boldsymbol{h}_v^t, \boldsymbol{h}_u^t, \boldsymbol{e}_{uv}\right), \\
&\boldsymbol{h}_v^{t+1} = U_t\left(\boldsymbol{h}_v^t, \boldsymbol{m}_v^{t+1}\right), \\
&\widehat{\boldsymbol{y}} = R\left(\left\{\boldsymbol{h}_v^T \mid v\in G\right\}\right).
\end{aligned} \tag{7}$$

在式 7 中, $\boldsymbol{m}$ 表示聚合的邻居信息, $\boldsymbol{h}$ 表示隐状态 (hidden state), $\boldsymbol{e}$ 表示节点 u, v 之间的关系. M, U, R 分别代表信息表示函数, 信息传递函数以及信息读出函数.

(iii) 未知关系推断给定 $\boldsymbol{y}$, 通过假设关系中在 (ii) 中求取的关系建模 f, 对 $\boldsymbol{y}, \boldsymbol{x}$ 之间的潜在关系 $\widehat{W}$ 进行推断. 这里侧重于潜在的关系进行推断和更新.

$$\widehat{W} = f(\mathbf{y}, \boldsymbol{x}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \mid \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}). \tag{8}$$

其中 $\boldsymbol{C}$ 表示任意宏观与微观要素集合及其关系网络.

这里, 推断的对象也可以从复杂系统单个关键要素拓展至多个要素及其构成的子系统.

在这里, (i) 侧重于对已知关系的表示; (ii) 侧重于对关系的建模和学习, (iii) 则侧重于对潜在关系的推断. 在关系推断原理与方法框架中, 网络模块化构成是层次模块化规律的重要基础. 网络模块既是生物网络关键要素的一种关系表示, 同时也可以作为未知关系推断的对象. 以网络模块为切入点, 可以实现宏, 微观生物网络关键要素的统一表示与关联分析, 进而实现网络关系的数学化与整体化.

关系推断方法原理的核心是把对象之间的关系作为一种新的信息引入进来, 本质上是引入了新的信息维度, 这和现有直接利用对象自身信息 (特征) 进行推断是有区别的. 同时, 关系推断也能体现”系统论” 的思想与理论, 系统论所强调” 整体大于局部之和”, 其本质也是因为系统考虑了局部要素之间的关系.

关系推断框架具有较大的包容性, 具体表现在:

1) 关系网络构建的对象可以涵盖多个层次, 包括表型, 组织, 细胞, 分子以及药物等, 形成多层次网络关系推断. 例如, 建立疾病表型 - 组织 - 分子的网络能够推测出组织特异性的模块, 实现高风险基因精准预测 [22]; 建立细胞 - 分子网络, 能够推断特定病变细胞的功能失调信号 [23]. 多层次网络关系推断还能突破多层次复杂生物系统解析壁垒, 构建具有全新结构的机器学习模型 [24], 例如建立能反映人类大脑运行机制的多层网络分析方法 [25]; 提出用于对具有多元和多层次信息的复杂数据进行建模和分析的数学框架 [26].

2) 关系表示与建模可涵盖多种关系类型, 包括基于先验知识的明确关系 (如蛋白相互作用关系) ; 基于预测分析与实验验证结合的关系 (如药物 - 靶标关系等) ; 基于关联分析的相似性关系 (如基因共表达关系) ; 基于表示学习的” 隐藏” 关系 (如基于 Network embedding, 图神经网络等方法提取的低维表示特征, 进而利用特征相似衡量关系的强度) 等 [21];

3) 关系推断可涵盖大多数推断方法, 是后者的一般形式. 例如, 关联分析本质上是一种基于共出现关系的特殊关系推断; 句法模式识别也是关系推断的一种简化形式; 而统计推断则可以被认为是一种基于样本概率分布的特殊关系推断.

3 基于生物网络的关系推断的代表性方法示例

关系推断原理为克服复杂生物系统解析难题提供了全新思路. 依据关系推断原理, 李梢课题组率先提出了 CIPHER [13] 等系列代表性算法, 实现疾病表型, 中西药物与分子之间关系的系统表示, 以及致病基因与药物靶点的高精度预测, 如图 3 所示.

以 CIPHER 算法为例, 其目的是从给定的表型出发, 从一组候选基因中寻找相关的干预基因. 依据” 层次模块化关系” 以及关系推断方法学框架, 该算法可分为以下几个主要步骤:

(i) 给定一个表型和一组候选基因, 从现有数据库中提取出: 标准化的表型及表型之间相似关系的集合; 疾病表型与已知致病基因的关系集合; 基因之间的关系网络 (蛋白质相互作用网络) .

(ii) 第一步: 将表型网络, 基因 - 表型网络与蛋白质相互作用网络整合成一个单一网络, 第二步: 从表型网络中提取给定表型与其余所有表型的相似度排序; 根据拓扑距离计算候选基因与其它所有疾病基因的邻近度并形成邻近度排序. 第三步, 利用线性回归模型计算邻近度与相似度排序的相关性, 作为一特性的分数分配给每个候选基因, 按分数排序可得所求的基因.

上述步骤可简化表示为:

$$R'(p,g) = \mathrm{Cor}_{p' \in NE(p,g)} \left(f\left(p,p'\right), f\left(g,p'\right) \right). \tag{9}$$

式中 p, g 分别为待推断关系的疾病表型与基因, p' 代表网络中疾病表型 p 与基因 g 的网络邻居

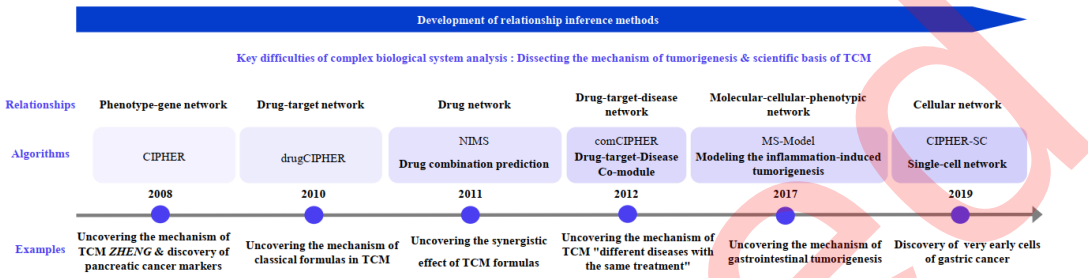

图 3　关系推断方法的发展 (李梢课题组)

Figure 3　Development of relationship inference-related algorithms (From Li Shao' s research group)

节点. CIPHER 算法利用最短路径对网络任意两点之间的关系进行了表示, 因此该算法具有全局预测能力. 此外, 我们通过在表型, 分子等不同层次网络中引入噪声信号, 对 CIPHER 算法的鲁棒性进行了系统的分析, 并与其他算法进行了比较. 结果发现, CIPHER 算法对致病基因预测的性能具有高度的鲁棒性, 在即使引入 30% 的噪声信号后, 预测精度依然可达到 0.35. 与包括基于贝叶斯模型 [27] 和基于随机游走模型 [28] 等非线性回归模型的算法进行比较, CIPHER 算法的预测性能高于这些算法最优值至少 43% [13], 提示基于线性回归模型的关系推断算法对生物网络噪声信号的高鲁棒性.

因此, CIPHER 算法建立了一个以多层次生物网络为基础, 从全局水平上定量推断致病基因的" 关系推断" 数学模型, 首次实现了疾病致病基因的全基因组预测, 其预测精度大幅提升至当时国际最高水平的 2 倍以上. CIPHER 方法已成功地预测出超过 5080 种疾病的致病基因谱, 并被广泛应用于胰腺癌, 肝癌等复杂疾病研究中, 取得了系列原创发现. 尤为重要的是, 利用 CIPHER 方法还首次实现全基因组层次的 400 余种中医表型相关基因预测, 揭示出中医经典寒, 热证的生物分子网络, 实现针对缺乏宏, 微观先验关联关系的复杂生物系统的关键要素推断.

借鉴关系推断的理论与方法学框架, 以及 CIPHER 算法所建立的数学模型, 李梢课题组还建立了系列方法, 包括前期的基于共出现和共表达构建分子网络的 LMMA 方法 [29], 预测中西药物靶标的 drugCIPHER 算法 [30], 实现大规模" 疾病 - 基因 - 药物" 共模块分析的 comCIPHER 算法 [31], 药物组合预测 NIMS 算法 [32], 融合多组学数据的致病基因突变预测算法 SPRING [33] 等. 其中, drugCIPHER 算法精度大幅超过国际最好方法的 6 倍, 实现 20 万余中西药物相关靶标预测. 系列方法被应用于解决中医药作用机理复杂, 现代适应症不清的瓶颈问题. 由此, 我们建立了一门新型交叉学科 - 中医药网络药理学 [34], 从系统层次和生物网络的整体角度出发, 解析疾病, 基因和药物之间的模块性关联规律, 为中医药现代化, 科学化提供新的思路与方法.

4　基于生物网络的关系推断方法的应用范例

4.1　肿瘤发生发展机制解析典型案例

目前, 关系推断方法在从分子, 细胞等多层次理解复杂疾病诊疗机制, 发现高可信度诊疗标志物方面取得了一些成功案例, 突显了关系推断方法在复杂疾病研究中的重要作用.

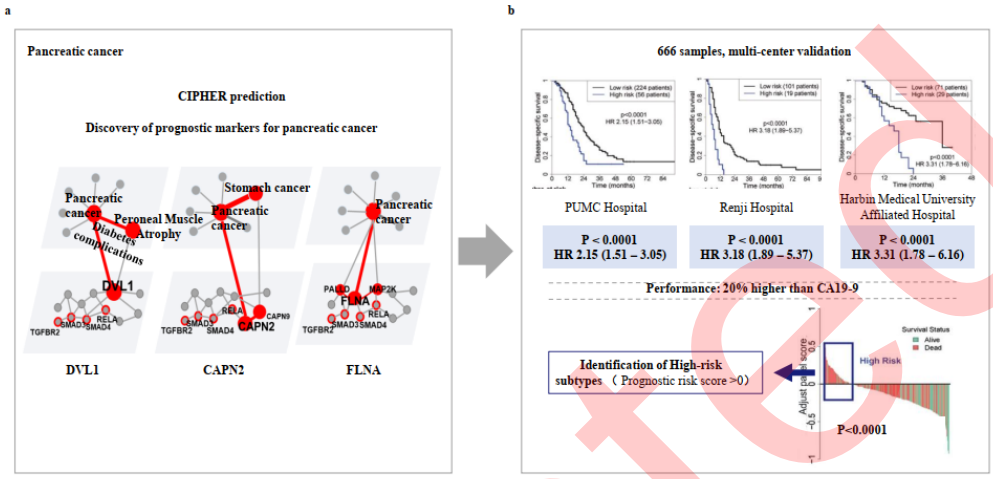

图 4　关系推断方法应用于胰腺癌研究的案例

Figure 4　An application case of relationship inference in pancreatic cancer

例如, 李梢课题组与协和医院赵玉沛院士团队通过紧密合作, 利用关系推断预测了胰腺癌高风险基因, 识别出胰腺癌预后和精准化疗标志物 [35] (图 4) . 首先利用 CIPHER 算法构建出胰腺癌相关的分子网络, 从分子网络中筛选出胰腺癌预后标志物网络模块, 优化出一组由五个关键节点组成的预后标志物组合. 然后, 经协和医院等多中心, 666 例患者临床验证, 发现该标志物组合具有显著的预后效果, 其判断预后效果显著优于现有标志物. 通过剖析算法预测原理, 发现胰腺癌与其他疾病 (如腓骨肌萎缩症, 胃癌等) 之间的表型相似" 关系" 的引入, 是从网络中识别出新标志物的关键 (图 4a) . 我们还利用关系推断方法来预测肝癌相关的关键模块和节点, 发现网络关键模块与肝癌发生, 预后显著相关, 预测结果与临床和实验具有较好的一致性 [36,37], 体现出关系推断方法在揭示复杂疾病诊疗机制方面的重要价值.

李梢课题组还将关系推断法方法拓展至细胞层次, 建立了细胞网络关系推断方法, 并将其应用解决我国高发的胃癌早诊难题. 如图 5 所示, 通过检测胃癌发生各阶段患者胃组织中近 10 万个细胞的基因表达, 并建立不同细胞之间定量关联关系的度量范式, 首次构建出胃炎癌转化单细胞网络; 进而, 利用细胞网络关系推断方法, 发现能标志胃炎癌转化" 癌变点" 的胃癌极早期细胞 (The very-early cells of gastric cancer) 及其标志物, 为胃癌防控前移提供全新的分期和靶标 [38]. 目前, 所发现的胃癌极早期细胞标志物已在全国 40 多家医院得到应用, 能显著前移胃癌早诊时间, 填补了胃癌极早诊断的空白. 这一案例也体现了关系推断方法在解决以肿瘤早诊标志物为代表的典型" 小样本" 问题上的重要价值, 为利用关系推断方法开展复杂疾病诊疗研究提供了全新的角度.

4.2　中药方剂作用机理解析典型案例

针对中药方剂复杂体系解析困难, 整体作用机制不清的瓶颈问题, 利用 drugCIPHER 等关系推断方法, 成功揭示白芍, 茯苓等扶正中药的生物学基础 [39]; 以寒热证和相关疾病分子网络为靶标, 发现血管新生调节中药方剂, 成分组合 [40,41]. 发现葛根芩连汤, 六味地黄丸等经典名方的新适应症 [42,43]. 其中, 葛根芩连汤治疗 2 型糖尿病新适应症, 获多中心随机双盲临床试验验证, 有力促进" 古方新

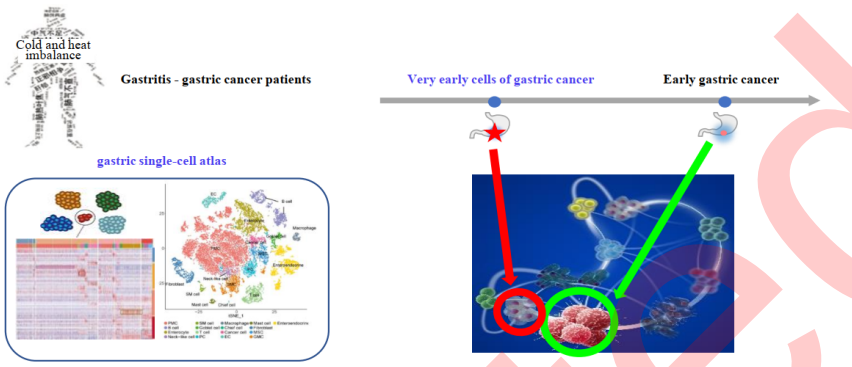

图 5　利用细胞网络关系推断, 发现胃癌极早期细胞

Figure 5　The discovery of the very-early cells of gastric cancer through relationship inference in cellular networks

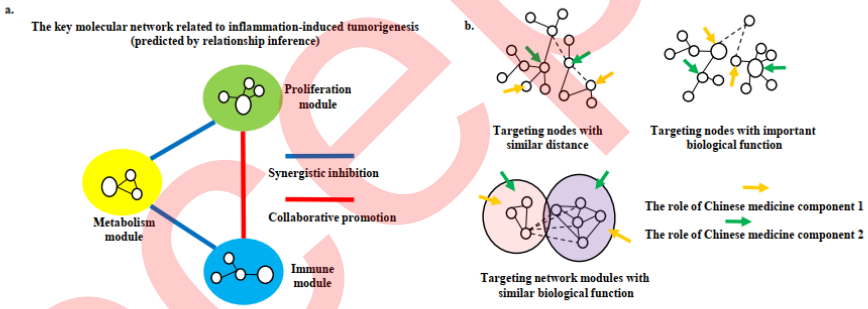

图 6　关系推断应用于炎癌转化及中药干预研究的案例

Figure 6　An application case of relationship inference in inflammation-induced tumorigenesis and traditional Chinese medicine intervention

用” [43].

如图 6 所示, 我们还利用关系推断方法预测了胃肠等炎癌转化的高风险基因, 构建出炎癌转化关键分子网络, 并结合联合敲低实验验证了关系推断方法所预测的高风险基因之间构成了具有协同作用的网络模块. 进而以炎癌转化关键子网的协同作用模块为干预靶标, 发现一些中药成分能够降低炎癌转化风险 [44]. 这不仅在一定程度上证明了关系推断方法的有效性, 也进一步突显了模块性规律在生物网络关系解析中的重要作用.

4.3　中医证候微观分子机制解析案例

关系推断的原理与方法还被应用于解决中医证候微观生物基础不清这类典型的” 小样本推断” 难题. 利用关系推断方法, 可以将疾病中, 西医宏观表型与微观生物分子作为” 整体” 用多层次生物网络进行表示, 并利用网络层次模块化关系, 将对疾病表型与生物分子之间的关系推断拓展到中医证候与生物分子之间的关系推断中, 从而实现中医证候分子机制的系统预测. 例如, 通过建立胃炎中西医表

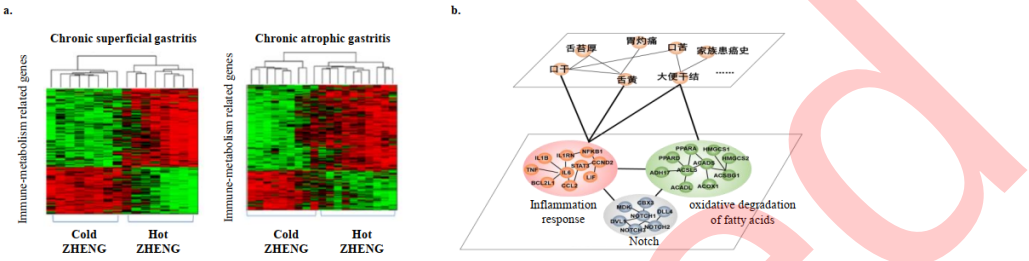

图 7　关系推断应用于胃炎寒热证研究的案例

Figure 7　An application case of relationship inference in Cold and Hot ZHENG related to gastritis

型与生物分子网络, 利用关系推断的原理与方法, 揭示出胃炎寒, 热证相关的代谢 - 免疫分子特征, 并得到临床组学验证 [45, 46](图 7a). 进而, 我们还识别出以寒热失衡, 代谢 - 免疫网络紊乱为特点的胃炎癌转化高风险亚型 (图 7b) , 促进胃炎癌转化的中医精准诊疗. 由此可见, 通过引入中医证候与疾病表型之间的关系, 并利用疾病表型与生物分子之间已知的关联, 可以实现对无已知基因的中医证候微观生物学基础的推断, 由此突破中医证候生物基础不清的难题, 为中西医结合的方法学研究提供了重要的切入点.

此外, 我们还基于关系推断方法, 提出了胃炎癌转化智能预警模型 [47], 发现了精准靶向胃癌极早期的药食同源中药等 [48], 建立了以" 智能早筛－极早诊断－精准早治" 为特点的胃癌三位一体防控模式, 该模式已在福州胃癌高发区人群推广, 成为" 国家慢性病综合防控示范区" 的创新亮点.

5　总结与展望

从复杂生物网络角度理解疾病的本质规律是当前科学研究的重大前沿, 为克服当前还原论医学研究模式的局限提供了重要突破口. 生物网络的高度复杂性决定了我们难以用传统的网络分析方法进行解析, 存在多层次信息难以融合以及" 小样本推断" 难题. " 模块性" 可能是复杂生物网络分析与干预的重要切入点. 本文揭示宏观疾病或药物与微观生物分子在生物网络中形成" 层次模块化关系", 提出了基于层次模块化关系的" 关系推断" 框架. 作为关系推断方法学实践, 我们提出的关系推断系列算法为疾病生物网络研究引入了一个新的维度, 在疾病基因预测, 中医药机理解析方面取得显著的效果, 并为网络靶标理论, 网络药理学等新兴学科提供了原理和方法上的支撑.

理解生物要素之间各种复杂" 关系" 是复杂生物网络研究的一个根本问题. 关系推断原理与方法框架的提出, 为突破复杂生物网络关系解析中多层次, 异质性信息融合难题以及" 小样本推断" 难题, 克服复杂生物系统难以建模与分析的重大挑战提供了一个解决方案, 也为从系统角度解析重大疾病生物机制, 创建系统干预手段提供了全新途径. 依据网络多层次相似性原理, 可从全局特征角度对疾病宏微观关联关系进行建模与分析, 实现多层次异质信息的融合; 借助疾病之间的相似性, 即便对致病基因完全未知的疾病也能进行推断, 实现无标签信息与先验知识的" 小样本" 推断. 尤为重要的是, 关系推断的原理与方法为克服中医药相关生物信息匮乏的难题, 揭示中医药分子机制也提供了系统解决方案. 中医药是中华民族的瑰宝, 然而长期以来, 缺少符合中医药整体特色的科研方法, 成为限制中医药创新发展的关键瓶颈. 中医" 整体" 诊疗思想所孕育的关系推断方法以及中医药网络药理学的建立与发展, 突破了这一关键瓶颈, 实现了中医药宏, 微观元素的关系表示与推断, 有望系统揭示中医药宏观

表型的微观生物学基础, 从而实现中医药研究理论与方法的原始创新, 推动中西医学的深入融合与未来发展, 为肿瘤等严重危害人类健康的重大疾病中西医防治带来新的希望.

网络的模块化构成原理是关系推断方法框架的基础. 这一点在将关系推断方法应用于胰腺癌, 肝癌, 胃癌等复杂疾病研究中也得到论证. 例如, 在胰腺癌预后标志物发现中, 我们得到的由五个具有显著预后价值的标志物在生物网络中就形成了密切相关的模块 (图 4a) . 剖析算法预测过程可以发现, 胰腺癌与腓骨肌萎缩症, 胃癌等疾病由于部分表型存在相似关系, 在表型网络中形成了模块. 同样地, 这些疾病已知的致病基因等在分子网络也具有模块效应, 因此, 从多层次生物网络中推断胰腺癌相关新标志物也就转变成对胰腺癌相关生物网络模块的推断. 网络模块化构成原理以及层次模块化规律具有一定的普适性, 上述推算思路也可以拓展应用多种复杂疾病相关生物网络分析.

关系推断法的效果取决于对" 关系" 的定义以及对先验知识的利用. 关系推断的结果还需要通过利用机理分析等先验知识进行筛选以及必要时利用实验进一步进行验证, 由于关系推断的结果已经大大缩小了筛选的范围, 因此可精准指导实验的开展. 同时, 关系推断方法也为其他领域的方法学研究提供了切入点. 例如, 在机器学习领域, 关系推断的原理与框架已逐渐被采用, 并展现出较大的发展潜力. 关系推断的策略在网络嵌入, 数据表示, 深度学习等领域也有较为广泛的应用. 在生物网络研究中, " 关系" 既包括宏观表型与微观分子的关联关系, 也包括疾病分子调控或药物干预的因果关系. 目前, 生物网络关系推断方法体系以关联分析为主, 反事实推理等因果推理的理论与方法应用还较少, 未来可深入挖掘蕴含在生物医学大数据中的因果关联, 进一步建立关联分析与因果推理相结合的生物网络关系推断方法, 为疾病发生发展的机理分析提供更为直接的计算证据. 总之, 关系推断对于中西医学, 人工智能以及大数据等多个领域研究都具有重要的意义, 具有广阔的应用前景.

致谢 感谢杨扩、侯思宇、王鑫等同学协助整理部分内容。

参考文献

1 Greene JA, Loscalzo J. Putting the patient back together - social medicine, network medicine, and the limits of reductionism. N Engl J Med. 2017;377(25):2493-2499.

2 Vidal M, Cusick ME, Barab á si AL. Interactome networks and human disease. Cell. 2011 18;144(6):986-98.

3 李梢. 中医证候与分子网络调节机制的可能关联. 中国科学技术协会首届学术年会 (1999 年, 杭州). 刊载于: 周光召主编. 面向 21 世纪的科技进步与社会经济发展, 第 1 版. 北京: 中国科学技术出版社.1999:442.

4 李梢, 王永炎, 季梁, 李衍达. 复杂系统意义下的中医药学及其案例研究. 系统仿真学报 2002;14(11):1429-1432.

5 Li S, Zhang Z, Wu L, et al. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. IET Syst Biol 2007;1(1):51-60.

6 李梢. 网络靶标：中药方剂网络药理学研究的一个切入点. 中国中药杂志 2011;36:2017-2020.

7 Liu YY, Slotine JJ, Barabási AL. Controllability of complex networks. Nature. 2011;473(7346):167-73.

8 Vinayagam A, Gibson TE, Lee HJ, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proc Natl Acad Sci U S A. 2016;113(18):4976-81.

9 Yan G, V é rtes PE, Towlson EK, et al. Network control principles predict neuron function in the Caenorhabditis elegans connectome. Nature. 2017;550(7677):519-523.

10 Park YP, Kellis M. CoCoA-diff: counterfactual inference for single-cell gene expression analysis. Genome Biol. 2021;22(1):228.

11 Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12(9):841-3.

12 Yan X, Liang A, Gomez J, et al. A novel pathway-based distance score enhances assessment of disease heterogeneity in gene expression. BMC Bioinformatics. 2017;18(1):309.

13 Wu X, Jiang R, Zhang MQ, et al. Network-based global inference of human disease genes. Mol Syst Biol. 2008;4:189.

14 Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol. 2015;7(3):214-30.

15 Santolini M, Barabási AL. Predicting perturbation patterns from the topology of biological networks. Proc Natl Acad Sci U S A. 2018;115(27):E6375-E6383.

16 Han P, Yang P, Zhao P, et al. GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 705-713.

17 Ruan P, Wang S. DiSNEP: a Disease-Specific gene Network Enhancement to improve Prioritizing candidate disease genes[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa241.

18 Kong W, Mou X, Hu X. Exploring matrix factorization techniques for significant genes identification of Alzheimer's disease microarray gene expression data. BMC Bioinformatics. 2011;12 Suppl 5(Suppl 5):S7.

19 Zhang M, Chen Y. Link prediction based on graph neural networks[J]. Advances in Neural Information Processing Systems, 2018, 31: 5165-5175.

20 Peng J, Guan J, Shang X. Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder. Front Genet. 2019;10:226.

21 Li X, Chen W, Chen Y, et al. Network embedding-based representation learning for single cell RNA-seq data. Nucleic Acids Res. 2017;45(19):e166.

22 Greene CS, Krishnan A, Wong AK, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47(6):569-76.

23 Sachs K, Perez O, Pe'er D, et al. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523-9.

24 Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks. Phys Rep. 2014;544(1):1-122.

25 De Domenico M. Multilayer modeling and analysis of human brain networks. Gigascience. 2017;6(5):1-8.

26 De Domenico M, Solé-Ribalta A, Cozzo E, et al. Mathematical formulation of multilayer networks[J]. Physical Review X, 2013, 3(4): 041022.

27 Lage K, Karlberg EO, St?rling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007;25(3):309-16.

28 Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24(5):537-44.

29 Li S, Wu L, Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics. 2006;22(17):2143-50.

30 Zhao S, Li S. Network-based relating pharmacological and genomic spaces for drug target identification. PLoS One. 2010;5(7):e11764.

31 Zhao S, Li S. A co-module approach for elucidating drug-disease associations and revealing their molecular basis. Bioinformatics. 2012;28(7):955-61.

32 Li S, Zhang B, Zhang N. Network target for screening synergistic drug combinations with application to traditional Chinese medicine. BMC Syst Biol. 2011;5 Suppl 1(Suppl 1):S10.

33 Wu J, Li Y, Jiang R. Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. PLoS Genet. 2014;10(3):e1004237.

34 Li S, Zhang B. Traditional Chinese medicine network pharmacology: theory, methodology and application. Chin J Nat Med. 2013;11(2):110-20.

35 Guo JC, Zhang P, Zhou L, et al. Prognostic and predictive value of a five-molecule panel in resected pancreatic ductal adenocarcinoma: A multicentre study. EBioMedicine. 2020;55:102767.

36 Su B, Luo T, Zhu J, et al. Interleukin-1 β /Iinterleukin-1 receptor-associated kinase 1 inflammatory signaling contributes to persistent Gankyrin activation during hepatocarcinogenesis. Hepatology. 2015;61(2):585-97.

37 Lin XM, Hu L, Gu J, et al. Choline kinase α mediates interactions between the epidermal growth factor receptor and mechanistic target of rapamycin complex 2 in hepatocellular carcinoma cells to promote drug resistance and xenograft tumor progression. Gastroenterology. 2017;152(5):1187-1202.

38 Zhang P, Yang M, Zhang Y, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Rep. 2019;27(6):1934-1947.e5.

39 Zheng J, Wu M, Wang H, et al. Network pharmacology to unveil the biological basis of health-strengthening herbal medicine in cancer treatment. Cancers (Basel). 2018;10(11):461.
40 Liao S, Han L, Zheng X, et al. Tanshinol borneol ester, a novel synthetic small molecule angiogenesis stimulator inspired by botanical formulations for angina pectoris. Br J Pharmacol. 2019;176(17):3143-3160.
41 Li S, Zhang B, Jiang D, et al. Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae. BMC Bioinformatics. 2010;11 Suppl 11(Suppl 11):S6.
42 Liang X, Li H, Li S. A novel network pharmacology approach to analyse traditional herbal formulae: the Liu-Wei-Di-Huang pill as a case study. Mol Biosyst. 2014;10(5):1014-22.
43 Li H, Zhao L, Zhang B, et al. A network pharmacology approach to determine active compounds and action mechanisms of ge-gen-qin-lian decoction for treatment of type 2 diabetes. Evid Based Complement Alternat Med. 2014;2014:495840.
44 Guo Y, Bao C, Ma D, et al. Network-based combinatorial CRISPR-Cas9 screens identify synergistic modules in human cells. ACS Synth Biol. 2019;8(3):482-490.
45 Li R, Ma T, Gu J, et al. Imbalanced network biomarkers for traditional Chinese medicine Syndrome in gastritis patients. Scientific Reports 2013;3:1543.
46 Wang X, Wu M, Lai X, et al. Network pharmacology to uncover the biological basis of spleen-qi-deficiency syndrome and herbal treatment. Oxidative Medicine and Cellular Longevity 2020: 2974268
47 Guo Y, Nie Q, MacLean A L, et al. Multiscale modeling of inflammation-induced tumorigenesis reveals competing oncogenic and oncoprotective roles for inflammation. Cancer Res. 2017;77(22):6429-6441.
48 Qi Q, Li R, Li HY, et al. Identification of the anti-tumor activity and mechanisms of nuciferine through a network pharmacology approach. Acta Pharmacol Sin. 2016;37(7):963-72.

Principle, method and application of relationship inference based on biological network

Shao Li[1*], Peng Zhang[1], Jin Gu[1], Rui Jiang[1] &
Yanda Li[1*]

1. *Institute for TCM-X, MOE Key Laboratory of Bioinformatics, Bioinformatics Division, BNRIST, Department of Automation, Tsinghua University, Beijing,* 100084, *China*
* Corresponding author. E-mail: shaoli@tsinghua.edu.cn, daulyd@tsinghua.edu.cn

Abstract In the era of big biomedical data, it has remained a common challenge for information science, Western medicine and traditional Chinese medicine (TCM) that systematically discovering key elements, including disease-causing genes and/or drug targets, followed by understanding the micro-level nature of macro-level phenotypes in the holistic fashion. The key to overcoming the challenge is how to solve the problem of multi-scale information fusion and that of the high-dimensionality, high-noise, and small-sample that exist in biomedical data, though in-depth understanding of the "relationship" nature of complex biological system (CBS), as biology is a typical complex system. Biological network, as the basis of complex biological systems, reflects the interrelationships of various biological molecules such as genes and gene products in the human body, as well as those between biological molecules and diseases and drugs at different levels. Biological networks have been widely used in biomedical sciences. Analysis of data. We started the research on the relationship between Chinese and Western medicine and CBN more than 20 years ago, and took the lead in proposing the hypothesis of "network target", and proceeding with method construction and application. In principle, this article uncovers a novel relationship named as 'multilevel modular relationship', between macro-level phenotypes and micro-level molecules based on CBN, followed by discussing CBN-based 'relationship inference'. It revealed that the macro-level emergence has local modularity at the micro level, and the more similar among macro-level phenotypes, the stronger the modular relationships among micro-level molecules (disease-causing genes or drug targets). Methodologically, we furtherly establish a general CBN-based computational framework for "relationship inference" to infer key elements from big biomedical data with the small number of positive samples, from a global perspective. It consists of three parts: 1) relationship network construction, 2) relationship representation and modeling, and 3) unknown relationships inferring, with aim to substantialization, mathematicization and integration of relationships. At the application level, the "relationship inference" framework have shown good performances in predicting disease-causing genes and drug targets, as well as identifying disease-related markers, and uncovering molecular mechanism related to TCM. Thus, this framework has provided a systematic solution for comprehensively understanding the micro-level nature of complex diseases and TCM. It has also provided important theoretical and methodological supports for some emerging disciplines, including network pharmacology.

Keywords Complex biological network, Modularity, Multilevel relationship, Relationship inference, Small-sample inference, Network pharmacology