中国科学: 信息科学 2019 年 第 49 卷 第 9 期: 1175–1185
SCIENTIA SINICA Informationis
论文

基于组织特异性和直接邻居相似度方法预测疾病 – 药物关系

鱼亮*, 赵晋

西安电子科技大学计算机科学与技术学院, 西安 710071
* 通信作者. E-mail: lyu@xidian.edu.cn

收稿日期: 2018–12–11; 修回日期: 2019–02–19; 接受日期: 2019–03–04; 网络出版日期: 2019–09–06

国家重点研发计划 (批准号: 2018YFC0910403)、国家自然科学基金面上项目 (批准号: 61672406) 和中央高校基本科研业务费 (批准号: JB180307) 资助项目

摘要 复杂疾病的致病机理一直是人类健康领域面临的重大难题之一, 通过传统的方法进行新药开发, 需要大量的时间与金钱, 已经满足不了人们的需求. 近几年来寻找已知药物新的治疗效果, 即药物重定位, 已经成为治疗更多疾病的一个有效途径. 目前组织特异性的研究已经取得一些成果, 但是传统的药物重定位方法很少考虑疾病的组织特异性. 本文提出基于组织特异性和直接邻居相似度方法预测药物的新适应症, 同时深入探讨考虑疾病的组织特异性对药物重定位研究的影响. 首先研究组织特异性的发展及其特点, 并提出基于组织特异性数据, 应用直接邻居的相似度进行药物重定位研究. 从数据库 DrugBank 中提取 11405 条已知药物 – 靶标关系, 并从人类孟德尔遗传数据库中获得 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌、卵巢癌) 及其致病基因数据, 利用 5 种癌症对应的组织特异性相互作用网络作为背景网络, 基于直接邻居距离度量方法构建 25 个组织特异性药物 – 疾病二部网络, 实验结果通过 CTD (comparative toxicogenomics database) 标准数据库进行验证. 结果表明, 基于组织特异性和直接邻居相似度度量标准会提高药物重定位研究的准确性, 为新药的体内和体外实验提供可靠候选集, 这也为药物重定位的研究提供了新的思路.

关键词 药物重定位, 组织特异性, 药物靶标, 致病基因, 直接邻居度量

1 引言

复杂疾病的研究一直是一个热门而且复杂的问题, 比如癌症、糖尿病, 以及心血管疾病往往是由于复杂的遗传因素和环境因素等共同导致的 [1]. 过去几十年中, 尽管基因组学和生命科学领域发展迅速, 但是新药的开发仍然非常耗时、花费巨大而且成功率很低, 导致新药的开发已经开始停滞 [2]. 据保

引用格式: 鱼亮, 赵晋. 基于组织特异性和直接邻居相似度方法预测疾病 – 药物关系. 中国科学: 信息科学, 2019, 49: 1175–1185, doi: 10.1360/N112018-00324
Yu L, Zhao J. Prediction of disease–drug relationships based on tissue specificity and direct neighbor similarity (in Chinese). Sci Sin Inform, 2019, 49: 1175–1185, doi: 10.1360/N112018-00324

守统计, 开发一种新药通常需要超过 15 年时间 [3] 并且花费 8 到 10 亿美元 [4], 所以当下有迫切的需求通过花费少量时间与金钱去开发新的药物.

面对这些挑战, 药物重定位 [5] 成为如今开发新药的重要方法之一, 它是一种发现现有药物新的临床适应症的策略, 它可以大大降低新药开发的风险和成本, 并且缩短药物发现和临床可用性之间的时间周期 [6]. 在 2013 年新上市的 84 个药物中有 20% 是重定位的药物 [7], 药物重定位在新药发现和精准医疗中起到了关键的作用. 然而, 早期药物新适应症的发现往往是偶然的, 比如砷作用于急性早幼粒细胞白血病 [8]、醋酸锌作用于威尔逊氏病 [9] 等, 都是通过临床实验偶尔发现的, 伴随着基因组学、生命科学、计算生物学的飞速发展以及各学科之间的融合交叉, 药物重定位正从依靠临床偶然发现慢慢转化为通过分析大规模生物数据进行精准的预测, 利用不同的数学模型以及计算的方法来提高预测药物与疾病之间关系的准确性.

随着生物领域技术的不断发展, 近几年来, 高通量基因测序技术的出现为研究人类基因组的秘密提供了强有力的帮助, 同时带来了许多有价值的基因表达数据, 比如药物作用于不同组织细胞系上的基因表达数据 [10]、不同癌症组织中的基因表达数据 [11]、正常组织中的基因表达数据 [12]、疾病状态下的基因突变数据 [12] 等. 众多类型的生物数据为药物重定位的研究提供了良好的基础, 而且药物重定位研究不仅仅局限于研究药物与药物之间的关系, 还可以研究药物与靶标基因、药物与疾病、药物与生物路径之间的关系等 [13].

根据不同的理论假设与数据, 研究人员提出了各种各样的方法来进行药物重定位研究, 主要的方法包括基于网络模型 [14~20]、基于药物的药理信息 [21~26]、基于化合物的化学结构 [27, 28]、基于药物的副作用 [29~32]、基于药物的靶标 [33, 34] 等, 这些方法利用不同类型的数据以及策略进行药物重定位研究, 使得研究人员能够在大大减少时间和金钱的情况下准确预测潜在的药物与疾病之间的关系.

另一方面, 组织特异性方面的研究也飞速发展. 组织特异性是由于某些组织特异表达基因的调控作用, 使得具有相同基因的细胞在不同的组织中往往呈现出差异较大的功能. 研究人员发现, 理解组织特异表达基因的功能和不同组织细胞系的作用, 对于疾病诊断和治疗以及复杂疾病的研究有很大帮助. Chen [35] 等从 TCGA (the cancer genome atlas) [11] 数据库中下载了 200 个肝癌组织的基因表达谱, 从 CCLE (cancer cell line encyclopedia) [36] 数据库中下载了超过 1000 个癌症细胞系的基因表达谱数据, 其中包含 25 个肝癌细胞系的基因表达谱数据. 通过研究发现, 肝癌组织的基因表达水平与肝癌细胞系中的基因表达水平很相近. Kosti 等 [37] 分析了来自 14 个不同组织类型中的 16561 个基因的表达水平, 以及与其对应的蛋白质功能, 发现相对于正常组织某些癌症致病基因与癌症组织之间的相关性更强. Guan 等 [38] 整合了基因组数据和组织特异性的基因表达数据, 构建了 107 个组织特异性的功能关系网络, 之后利用这些组织特异性的网络预测基因对应的表型, 并且证明了整合组织特异性可以提高预测的准确率.

目前全基因组关联研究已经确定了成千上万个常见疾病的致病基因, 但是, 大多数的复杂疾病的致病机制仍然是比较模糊的. Lonsdale 等 [39] 为了深入了解不同疾病的发病机理, 建立了一个数据库 GTEx (genotype-tissue expression) 专门研究人类不同组织中的遗传变异和基因表达之间的关系. Pierson 等 [40] 利用 GTEx 中的数据建立了 35 个人类组织特异的基因共表达网络, 发现具有组织特异性功能的转录因子, 通常在共表达网络中处于中心节点位置, 且与具有组织特异性功能的基因相连. Guo 等 [41] 通过整合大规模的转录因子数据和基因表达数据在 13 个不同组织中建立组织特异性的基因调控网络, 通过比较这些调控网络, 发现组织特异性的调控因子比在多个组织中表达的调控因子能够调控更多的基因, 并且由组织特异性的调控因子调控的生物过程与组织的功能相关性很大. 因此研究组织特异性调控网络, 可以更好地帮助我们了解复杂疾病的分子机制以及确定新的疾病致病基因.

表 1 药物数据格式

Table 1 Drug data format

DrugBank ID	Drug name	Entrez gene ID
DB00001	Lepirudin	2147
DB00007	Leuprolide	2798

目前对于组织特异性的研究虽然已经取得了一定的成功, 包括对于复杂疾病致病机理的研究, 以及不同组织中基因表达数据的挖掘处理, 并且证明了许多疾病与其对应的组织之间的关系非常紧密 [42]. 但是, 很少有算法利用组织特异性数据进行药物重定位研究. 本文基于组织特异性, 将直接邻居的方法应用于预测药物与疾病的潜在关系. 基于 3 种数据: (1) 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌和卵巢癌) 对应的组织特异性相互作用网络数据; (2) 11405 对药物 – 靶标数据; (3) 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌和卵巢癌) 及其致病基因数据. 构建了 25 个组织特异性药物 – 疾病二部网络, 来预测 5 种癌症的潜在治疗药物, 并研究组织特异性网络对预测癌症新药的影响. 通过标准数据库和文献验证说明, 针对某种疾病并且整合该疾病对应的组织信息进行药物重定位研究, 大大提高了预测结果准确率.

2 基于组织特异性和直接邻居相似度预测疾病 – 药物关系

2.1 组织特异性蛋白质网络、药物 – 靶标, 以及疾病 – 基因数据

本研究工作使用了 3 种数据: 药物 – 靶标关系数据、疾病与致病基因的关系数据和组织特异性蛋白质相互作用网络数据.

组织特异性蛋白质相互作用网络数据. 蛋白质相互作用网络在本研究中起到非常重要的作用, 其可靠性直接决定了研究结果的精确度. 2015 年, Greene 等 [43] 利用数据驱动的 Bayes 方法整合了成千上万的跨组织和疾病状态的实验数据, 构建了 144 个人类组织的蛋白质功能相互作用网络, 为识别基因的跨组织功能作用变化, 以及研究复杂疾病与组织之间的关系提供了强有力的帮助, 并且创建了网站 GIANT (genome-scale integrated analysis of gene networks in tissues)[1)]. 从 GIANT 上可以下载所有组织特异的蛋白质相互作用网络数据, 该数据库共包含 130 多个不同组织的蛋白质相互作用网络, 至今已经被引用 295 次.

药物 – 靶标关系数据. 该数据从 DrugBank [44] 数据库中下载, 是目前最新的 FDA 认证数据. 对下载到的数据进行预处理, 如果药物没有靶标, 则将该药物进行过滤, 最终得到 1679 个药物和 1634 个靶标基因, 它们之间共有 11405 对药物 – 靶标关系. 药物采用 DrugBank 数据库中的 ID 作为唯一标识, 靶标基因用 Entrez gene ID 唯一标识. Entrez gene ID 是指来自于 NCBI (national center for biotechnology information) 旗下的 Entrez gene 数据库所使用的编号, 对应于染色体上一个基因位置 (gene location). 药物和靶标的数据格式示例如表 1 所示.

疾病 – 基因数据. 疾病与致病基因的关系数据从 OMIM (online mendelian inheritance in man) [45] 数据库中得到. 该数据库主要关注人类基因变异和表型性状之间的关系, 着眼于可遗传的或遗传性的基因疾病, 包括文本信息和相关参考信息、序列纪录、图谱和相关其他数据库, 一直持续更新. 数据库中的每一条记录, 有一个唯一由 6 位数字组成的 OMIM 编号 (OMIM ID), 不同数字开头的编号含义

1) http://giant.princeton.edu/.

表 2 疾病数据格式

Table 2 Disease data format

OMIM ID	Cancer name	No. disease-causing genes
114480	Breast cancer	31
114500	Colon cancer	33
114550	Liver cancer	18
211980	Lung cancer	24
167000	Ovarian cancer	10

表 3 组织特异性蛋白质相互作用网络数据

Table 3 Tissue-specific protein interaction network data

Cancer name	Tissue name	No. nodes	No. edges
Breast cancer	Breast	3318	30000
Colon cancer	Colon	3909	30000
Liver cancer	Liver	3360	30000
Lung cancer	Lung	3648	30000
Ovarian cancer	Ovary	3811	30000

不同, 详见网址[2)]. 本研究中, 我们选择了 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌和卵巢癌) 作为研究对象, 下载了其对应的致病基因数据. 选择这 5 种癌症的主要原因是, 目前针对它们的研究相对比较多, 数据比较全面. 其中, 癌症名称用 OMIM ID 唯一标识, 致病基因用 Entrez gene ID 唯一标识, 癌症数据以及对应致病基因的数量如表 2 所示.

5 种癌症对应的组织特异性蛋白质相互作用网络数据, 来自于 GIANT 数据库中下载的最新数据. 网络规模较大、边均有权值, 并且不同组织特异性网络建立所整合的数据规模不一致. 比如, 在不同网络中用相同的阈值过滤边, 发现乳腺组织蛋白质相互作用网络中边的数量几乎是其他癌症的两倍. 这是因为有关乳腺组织的研究较多, 因而相关的生物数据信息比较丰富. 为了使得 5 种组织特异性蛋白质相互作用网络规模比较相当, 我们选择利用边的数量来控制网络的规模, 获得高可信的子网络. 因为 GIANT 数据库中的边权值越大, 代表节点之间的相关性越可靠. 因此, 最终我们选定每个组织特异性网络中, 边权值排在前 30000 的边, 以及它们相关的基因节点作为下一步研究的对象. 这样获得具有相同边数, 但是结构各不相同的 5 个癌症组织对应的组织特异性蛋白质相互作用网络. 具体的 5 个子网络边数, 以及每个子网络中所包含的基因节点个数如表 3 所示.

2.2 直接邻居相似度量方法

利用网络模型计算药物 – 疾病之间的关系, 比较常见的方法是将蛋白质相互作用网络作为背景网络, 将药物的靶标基因和疾病的致病基因分别投影到蛋白质相互作用网络中, 形成药物靶标基因模块和疾病致病基因模块, 然后利用网络拓扑结构计算两个模块之间的距离, 并将该距离作为药物 – 疾病之间的相似性. 常用的计算网络中模块距离的方法包括: 直接邻居算法、最短路径算法、随机游走算法 [46], 以及标签传播算法 [47] 等. Wu 等 [48] 分别使用了直接邻居与最短路径两种方法预测人类疾病与致病基因之间的关系, 通过实验结果分析发现, 利用直接邻居构建疾病 – 基因关系的方法比最短路

2) https://omim.org/.

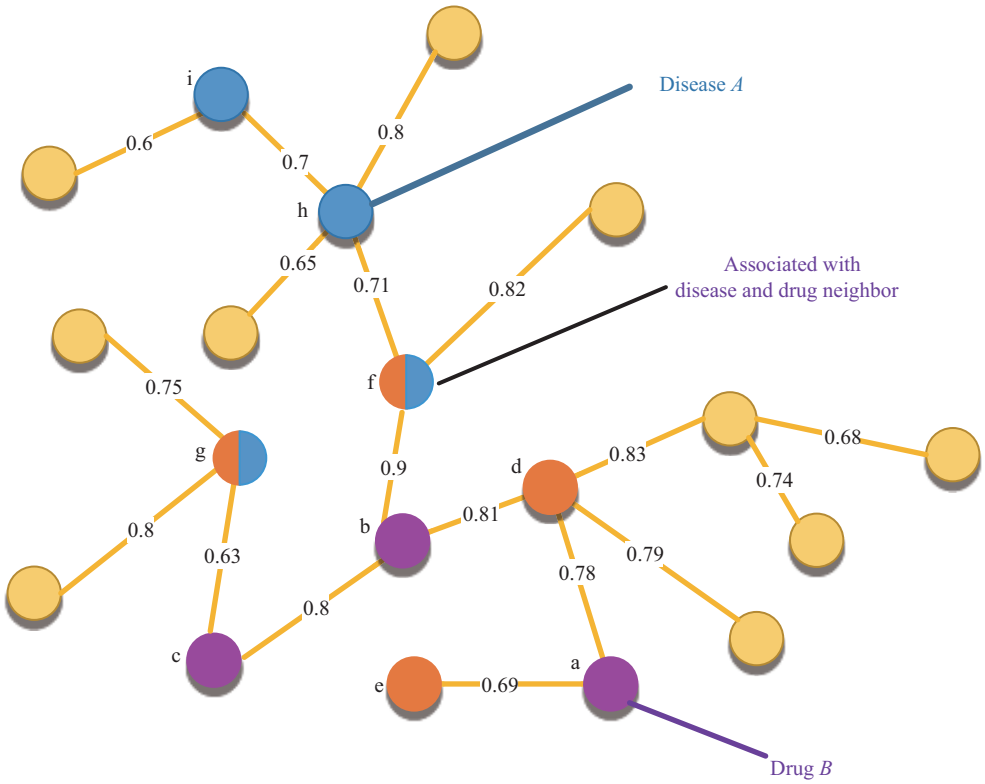

图 1 (网络版彩图) 基于直接邻居计算药物 – 疾病相似性的实例

Figure 1 (Color online) Example of calculating drug-disease similarity based on direct neighbors

径的方法准确率高. 因此, 在本研究我们采用直接邻居度量方法来建立药物与疾病之间的关系, 进而研究组织特异性对药物重定位的影响.

图 1 给出一个利用直接邻居度量方法 [48] 计算疾病 A 与药物 B 关系的实例. 图 1 中所有节点代表蛋白质相互作用网络中的基因, 其中蓝色节点集 $\{\mathrm{f},\mathrm{g},\mathrm{h},\mathrm{i}\}$ 为疾病 A 的致病基因集合, 紫色节点集 $\{\mathrm{a},\mathrm{b},\mathrm{c}\}$ 为药物 B 的靶标基因集合, 橙色节点集 $\{\mathrm{d},\mathrm{e},\mathrm{f},\mathrm{g}\}$ 为药物靶标基因在背景网络中的直接邻居基因集合, g 和 f 节点具有橙色和蓝色两种颜色, 表示药物靶标及其直接邻居基因构成的节点集合 $\{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e},\mathrm{f},\mathrm{g}\}$ 与疾病基因集合 $\{\mathrm{f},\mathrm{g},\mathrm{h},\mathrm{i}\}$ 共有的节点. 最终疾病 A 与药物 B 之间的相似性 $S(A,B)$ 的计算公式如下所示:

$$S(A,B)=\sum_{i=1}^{n}k_i, \tag{1}$$

其中 n 代表药物相关的基因集合与疾病致病基因集合重叠的基因个数, 例如在图 1 中 $\{\mathrm{f},\mathrm{g}\}$ 是疾病 A 的致病基因集合与药物 B 的靶标及其直接邻居的重叠基因集, 所以 $n=2$; k_i 代表药物靶标基因与重叠基因之间的权值, 如图 1 所示, 边 (b, f) 上的权值为 0.63, 边 (c, g) 上的权值为 0.9, 所以 $S(A,B)=0.63+0.9=1.53$, 即疾病 A 与药物 B 之间的相关性大小为 1.53. 如果药物的靶标基因直接与疾病基因重叠, 则对应的边权值 $k_i=1$. 由于我们所使用的蛋白质相互作用网络中, 边的权重与节点之间的关系强弱正相关, 所以 $S(A,B)$ 的值越大代表疾病 A 与药物 B 之间的关系越紧密.

表 4　5 个基于直接邻居度量的药物 – 疾病网络信息

Table 4　Five drug-disease network information based on direct neighbor metrics

Cancer name	Tissue name	No. drug nodes	No. edges
Breast cancer	Breast	281	596
Colon cancer	Colon	298	749
Liver cancer	Liver	283	557
Lung cancer	Lung	281	596
Ovarian cancer	Ovary	305	625

2.3　基于直接邻居相似度量预测疾病 – 药物关系

基于直接邻居相似度量预测疾病 – 药物关系, 主要分为以下几个步骤:

首先, 从 GIANT 数据库中分别下载 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌和卵巢癌) 对应的 5 个组织特异性蛋白质相互作用网络, 将它们作为背景网络;

接着, 将 5 种癌症对应的致病基因投影到 5 个组织特异性蛋白质网络中, 即每种癌症的致病基因都会投影到 5 个网络中;

再者, 将 1679 个药物对应的靶标基因, 投影到 5 个蛋白质相互作用网络中, 并获得这些靶标基因的直接邻居节点;

最后, 根据直接邻居相似度度量方法, 在每一个组织特异性蛋白质网络中, 分别计算 1679 个药物与癌症之间的相关性, 即对于每种癌症, 分别以 5 种网络作为背景网络, 计算该癌症基于不同组织特异性网络获得的潜在药物, 目的是研究组织特异性对药物重定位的影响, 这样可以得到 25 个药物 – 疾病二部网络.

表 4 给出了每种癌症在其对应组织中进行计算后, 得到药物 – 疾病二部网络信息. 因为选择的 5 个蛋白质相互作用网络规模相当, 所以得到的二部网络规模也相差不大.

3　预测结果分析

基于计算的结果, 首先分析在不同组织特异性蛋白质相互作用网络中, 每种癌症得到的候选药物结果. 以乳腺癌为例, 在 5 个蛋白质网络中, 分别得到与乳腺癌相关的 5 个候选药物列表, 将每个列表中的候选药物按照相似性的值进行降序排列. 我们将对 5 个药物列表中的前 10 个药物进行具体分析, 这里选择前 10 个药物的原因是, 预测结果排名越靠前的药物, 与疾病的关系准确率越高, 具体的结果如表 5 所示.

我们进一步通过比较毒物遗传学数据库 (comparative toxicogenomics database, CTD) [49] 对预测结果进行验证. CTD 是一个强大、公开可用的数据库, 旨在提高人们对环境变化是如何影响人类健康的理解. 它提供了许多可靠的信息, 包括化合物 – 基因关系、化合物 – 疾病关系, 以及基因 – 疾病关系, 这些信息集成生物功能网络数据, 可以帮助人们理解复杂疾病的发病机理. CTD 数据库中记录了所有在文献中报道过的药物 – 疾病关系, 在表 5 中药物的 DrugBank ID 后面带 * 表示该药物 – 疾病关系可以在 CTD 中查找到, 并且其 inference score 大于 10, 这里的 inference score 反映了 CTD 化合物 – 基因 – 疾病关系网络与类似的无标度随机网络之间的相似程度. 分数越高, 说明 CTD 化合物 – 基因 – 疾病关系网络越可靠. 带 ** 的表示该药物 – 疾病关系在 CTD 中被标注为 “M” (表示化合物与疾病相关或可能在疾病的病因学中起作用), “T” (表示化合物在疾病中具有已知或潜在治疗作用),

表 5 5 个基于直接邻居度量的药物 − 疾病网络信息

Table 5 Five drug-disease network information based on direct neighbor metrics

Tissue name	Breast	Colon	Liver	Lung	Ovary
DrugBank ID	DB00201**	DB00201**	DB00997**	DB00201**	DB00128
	DB00563**	DB00563**	DB04967*	DB04967*	DB00563**
	DB00997**	DB00642**	DB08818*	DB00997**	DB00642**
	DB00675**	DB00945**	DB00201**	DB00675**	DB01183*
	DB08818*	DB00997**	DB00642**	DB01108	DB08818*
	DB00242*	DB00432*	DB00563**	DB08818*	DB01708*
	DB00642**	DB04967*	DB01183*	DB00242*	DB09074**
	DB04967*	DB00440	DB00218	DB00642**	DB00675**
	DB00128	DB06813	DB00276**	DB00563**	DB00171
	DB01183*	DB00218	DB00380	DB00128	DB00440

表 6 5 种癌症在 5 种组织中的候选药物对比 CTD 数据库的统计结果

Table 6 Comparison of drug candidates in five tissues versus CTD database for five cancers

Tissue name	Breast cancer	Colon cancer	Liver cancer	Lung cancer	Ovarian cancer
Breast	9	5	6	8	8
Colon	7	8	4	8	5
Liver	8	5	6	9	7
Lung	8	6	7	8	6
Ovary	7	8	4	9	6

或者 “M|T”.

通过表 5 可以看出乳腺癌在乳腺组织中排名前 10 的预测结果中, 有 9 个在 CTD 数据库中有记录, 其中有 5 个是非常可信的关系, 即药物的 ID 号后带了 **. 以药物咖啡因 (DrugID: DB00201) 和氨甲叶酸 (DrugID: DB00563) 为例. 在 CTD 数据库中, 咖啡因 (caffeine) 被标注为 M, 即标记 (marker) 或机制 (mechanism). 研究者经过实验研究发现 [50,51], 含较高咖啡因的咖啡摄入量可能与绝经后乳腺癌的风险较低有关. 对于氨甲叶酸 (Methotrexate), 它在 CTD 数据库中被标记为 T, 即治疗 (therapeutic). 目前已有大量研究及临床实验 [52~54] 发现, 氨甲叶酸经常与其他化疗药物联合使用进行乳腺癌的治疗. 另外, 我们还发现, 对于药物天冬氨酸 (DrugID: DB00128), 虽然没有在 CTD 数据中有记录, 但是也已经有文献说明 [55], 低循环天冬氨酸是人类乳腺癌的一个重要代谢特征. 而基于其他 4 个组织网络, 预测结果的前 10 个中至少有 2 个是未出现在 CTD 标准数据中. 由此可以说明, 利用乳腺组织蛋白质相互作用网络进行乳腺癌的药物重定位研究, 可以在一定程度上提高预测结果的准确性. 对于所有的 5 种癌症, 我们分别计算了在 5 种不同组织中得到的潜在药物, 并如表 5 一样, 对同一癌症, 分别比较了在 5 种不同组织中, 预测得到的药物是否在 CTD 标准数据库中有记录, 统计结果展示在表 6 中. 表 6 表示每种癌症在不同组织中, 预测得到的前 10 个得分最高的药物, 与 CTD 标准数据库对比后, 在 CTD 数据库中有记录的药物数量. 例如在 “结肠” 组织中, “结肠癌” 对应的数字是 “8”, 那么说明在结肠组织中, 预测结肠癌相关的药物, 排在前 10 的药物中, 有 8 个药物在 CTD 标准数据库有记录表明与结肠癌相关.

从表 6 可以看出, 对角线上的数字表示癌症在最相关的组织中, 预测得到的相关药物在 CTD 数

据库中有记录的数量, 发现乳腺组织和结肠组织中对应药物数量都排第 1 位, 肺组织、肝组织和卵巢组织中对应预测的药物数量尽管没有排名第一, 但也排名靠前. 关于肺组织、肝组织和卵巢组织对应数据, 出现这样结果的原因可能是因为不同癌症的致病基因数量不同. 据表 2, 我们知道乳腺癌、结肠癌、肺癌、卵巢癌和肝癌对应的致病基因数量分别是: 31, 33, 24, 10, 18. 很显然乳腺癌、结肠癌比肺癌、卵巢癌和肝癌的致病基因多了很多. 这也可能是肺癌、卵巢癌和肝癌结果与其他两种癌症有差别的原因之一. 乳腺癌、结肠癌和肺癌的前 10 个预测药物中, 至少有 8 个都在 CTD 标准数据中有记录, 也就是预测准确率达到了 80% 以上, 而对于肝癌和卵巢癌来讲, 尽管预测的已知疾病 – 药物关系较少, 也都达到了 6 个, 这也进一步说明, 基于组织特异性网络和直接邻居距离度量方法进行药物重定位研究, 准确性非常高, 可以为疾病的新药预测提供可靠候选. 另外, 从表 6 还可以看出, 对于相似的乳腺和肝组织 [56], 它们具有相似的预测结果. 对于这 5 种癌症来讲, 除了在 CTD 标准数据库中已经有记录的药物之外, 其余的药物很可能就是潜在的治疗疾病的新药, 可以作为进一步实验的候选药物, 这为新药的研制和发现节省了大量的时间和经费. 这些结果说明在疾病对应组织中进行药物重定位研究的新思路可靠、可行.

4 结束语

本文以组织特异性蛋白质网络为背景网络, 将基于直接邻居的距离度量应用于药物重定位研究, 一方面说明, 基于直接邻居距离度量能够很好地预测药物与疾病之间的关联关系; 另一方面也说明, 考虑疾病的组织特异性能够大大提高疾病 – 药物关系预测结果的准确性. 我们集成了 GIANT 中的组织特异性蛋白质相互作用网络、DrugBank 中的药物与靶标关系, 以及数据库 OMIM 中得到的疾病与其致病基因之间的关系数据, 并利用基于直接邻居的距离度量方法构建了 5 种癌症 (乳腺癌、结肠癌、肝癌、肺癌、卵巢癌) 在不同组织中对应的癌症 – 药物二部网络, 基于构建的二部网络预测这 5 种癌症潜在的治疗药物, 选取 5 种癌症在不同组织中得到的前 10 个潜在药物进行了 CTD 标准数据库验证. 验证结果发现, 应用与疾病对应的组织特异性蛋白质相互作用网络, 进行潜在药物预测准确性非常高, 在前 10 个预测结果中, 乳腺癌、结肠癌和肺癌的准确率达到了 80% 以上, 那么新预测出来的一种或者两种药物, 很有可能就是治疗该癌症的潜在药物, 这就为新药的发现提供了高可信的候选参考, 进一步说明了组织特异性以及基于直接邻居距离度量方法的可靠性和高效性.

未来研究工作将对本文的研究结果进行深入分析, 进一步进行细胞系、小鼠等生物实验验证, 并将本文中提到的方法推广到其他疾病中, 希望能为药物重定位研究提供新的研究思路.

参考文献

1 Schork N J. Genetics of complex disease: approaches, problems, and solutions. Am J Respir Crit Care Med, 1997, 156: 103–109

2 Booth B, Zemmel R. Prospects for productivity. Nat Rev Drug Discov, 2004, 3: 451–456

3 DiMasi J A, Bryant N R, Lasagna L. New drug development in the United States from 1963 to 1990. Clin Pharmacol Therory, 1991, 50: 471–486

4 Adams C P, Brantner V V. Estimating the cost of new drug development: is it really 802 million dollars? Health Aff, 2006, 25: 420–428

5 Ashburn T T, Thor K B. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov, 2004, 3: 673–683

6 Benkimoun P. French health minister announces reform of drug regulation. BMJ, 2011, 342: 360

7 Graul A, Cruces E, Stringer M. The year's new drugs & biologics, 2013: Part I. Drugs Today, 2014, 50: 51–100

8 Soignet S L, Maslak P, Wang Z G, et al. Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. N Engl J Med, 1998, 339: 1341–1348
9 Scheindlin S. Rare diseases, orphan drugs, and orphaned patients. Mol Interv, 2006, 6: 186–191
10 Subramanian A, Narayan R, Corsello S M, et al. A next generation connectivity map: L1000 platform and the first 1000000 profiles. Cell, 2017, 171: 1437–1452
11 Hutter C, Zenklusen J C. The cancer genome atlas: creating lasting value beyond its data. Cell, 2018, 173: 283–285
12 Barrett T, Wilhite S E, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res, 2012, 41: 991–995
13 Wu Z K, Wang Y, Chen L N. Network-based drug repositioning. Mol Biosyst, 2013, 9: 1268–1281
14 Yu L, Huang J B, Ma Z X, et al. Inferring drug-disease associations based on known protein complexes. BMC Med Genomics, 2015, 8: 13
15 Yu L, Ma X K, Zhang L, et al. Prediction of new drug indications based on clinical data and network modularity. Sci Rep, 2016, 6: 32530
16 Iorio F, Saez-Rodriguez J, Bernardo D. Network based elucidation of drug response: from modulators to targets. BMC Syst Biol, 2013, 7: 139
17 Li J, Lu Z Y. Pathway-based drug repositioning using causal inference. BMC Bioinform, 2013, 14: 3
18 Zhao H, Jin G X, Cui K M, et al. Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. Cancer Res, 2013, 73: 6149–6163
19 Cui H, Zhang M H, Yang Q M, et al. The prediction of drug-disease correlation based on gene expression data. Biomed Res Int, 2018, 2018: 1–6
20 Liang X J, Zhang P F, Yan L, et al. LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. Bioinformatics, 2017, 5: 770
21 Yu L, Su R D, Wang B B, et al. Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. IEEE/ACM Trans Comput Biol Bioinf, 2017, 14: 966–977
22 Yu L, Wang B B, Ma X K, et al. The extraction of drug-disease correlations based on module distance in incomplete human interactome. BMC Syst Biol, 2016, 10: 111
23 Bisgin H, Liu Z C, Kelly R, et al. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC Bioinform, 2012, 13: 6
24 An S M, Ding Q P, Li L. Stem cell signaling as a target for novel drug discovery: recent progress in the WNT and Hedgehog pathways. Acta Pharmacol Sin, 2013, 34: 777–783
25 Zhang W, Yue X, Huang F, et al. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods, 2018, 145: 51–59
26 Zhang W, Yue X, Lin W R, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinform, 2018, 19: 233
27 Novick P A, Ortiz O F, Poelman J, et al. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. PLoS ONE, 2013, 8: 79568
28 Wang Y C, Chen S L, Deng N Y, et al. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS ONE, 2013, 8: 78518
29 Yu L, Zhao J, Gao L. Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. Artif Intell Med, 2017, 77: 53–63
30 Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. PLoS ONE, 2011, 6: 28025
31 Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLoS ONE, 2014, 9: 87864
32 Bisgin H, Liu Z C, Fang H, et al. A phenome-guided drug repositioning through a latent variable model. BMC Bioinform, 2014, 15: 267
33 Yu L, Zhao J, Gao L. Predicting potential drugs for breast cancer based on miRNA and tissue specificity. Int J Biol Sci, 2018, 14: 971–982
34 Swamidass S J. Mining small-molecule screens to repurpose drugs. Brief Bioinform, 2011, 12: 327–335
35 Chen B, Sirota M, Fan-Minogue H, et al. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Med Genomics, 2015, 8: 5

36 Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012, 483: 603–607
37 Kosti I, Jain N, Aran D, et al. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci Rep, 2016, 6: 24799
38 Guan Y F, Gorenshteyn D, Burmeister M, et al. Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS Comput Biol, 2012, 8: 1002694
39 Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. Nat Genet, 2013, 45: 580–585
40 Pierson E, Koller D, Battle A, et al. Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput Biol, 2015, 11: 1004220
41 Guo W L, Zhu L, Deng S P, et al. Understanding tissue-specificity with human tissue-specific regulatory networks. Sci China Inf Sci, 2016, 59: 070105
42 Ni J C, Koyuturk M, Tong H H, et al. Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. BMC Bioinform, 2016, 17: 453
43 Greene C S, Krishnan A, Wong A K, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet, 2015, 47: 569–576
44 Wishart D S, Knox C, Guo A C, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res, 2006, 34: 668–672
45 Hamosh A, Scott A F, Amberger J S, et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2004, 33: 514–517
46 Chen X, Liu M X, Yan G Y. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst, 2012, 8: 1970–1978
47 Chen X, Zhang D H, You Z H. A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. J Transl Med, 2018, 16: 348
48 Wu X, Jiang R, Zhang M Q, et al. Network-based global inference of human disease genes. Mol Syst Biol, 2008, 4: 189
49 Davis A P, Grondin C J, Johnson R J, et al. The comparative toxicogenomics database: update 2017. Nucleic Acids Res, 2017, 45: 972–978
50 Bhoo-Pathy N, Peeters P H, Uiterwaal C S, et al. Coffee and tea consumption and risk of pre- and postmenopausal breast cancer in the European prospective investigation into cancer and nutrition (EPIC) cohort study. Breast Cancer Res, 2015, 17: 15
51 Ganmaa D, Willett W C, Li T Y, et al. Coffee, tea, caffeine and risk of breast cancer: a 22-year follow-up. Int J Cancer, 2008, 122: 2071–2076
52 Schilsky R L. Methotrexate: an effective agent for treating cancer and building careers. The polyglutamate era. Oncologist, 1996, 1: 244–247
53 Kumaki N, Okamatsu C, Tokuda Y, et al. Breast cancer in patients of rheumatoid arthritis with methotrexate therapy mimicking histopathological changes after neoadjuvant chemotherapy. Tokai J Exp Clin Med, 2017, 42: 104–108
54 Colleoni M, Rocca A, Sandri M T, et al. Low-dose oral methotrexate and cyclophosphamide in metastatic breast cancer: antitumor activity and correlation with vascular endothelial growth factor levels. Ann Oncol, 2002, 13: 73–80
55 Xie G X, Zhou B S, Zhao A H, et al. Lowered circulating aspartate is a metabolic feature of human breast cancer. Oncotarget, 2015, 6: 33369–33381
56 Uhlen M, Fagerberg L, Hallstrom B M, et al. Tissue-based map of the human proteome. Science, 2015, 347: 1260419

Prediction of disease–drug relationships based on tissue specificity and direct neighbor similarity

Liang YU* & Jin ZHAO

School of Computer Science and Technology, Xidian University, Xi'an 710071*, China*
* Corresponding author. E-mail: lyu@xidian.edu.cn

Abstract The pathogenesis of complex diseases is a major problem in the field of human health. The development of new drugs through traditional methods requires considerable time and money, which has not met people's actual requirements. Recently, identifying new therapeutic effects of known drugs via drug repositioning has become an effective way to treat numerous diseases. At present, tissue-specific research has achieved some success; however, traditional drug repositioning methods rarely consider the tissue specificity of the disease. To explore the influence of tissue specificity on drug repositioning studies, this study explores the development of tissue specificity and its characteristics and proposes using direct neighbor similarity in drug repositioning based on tissue-specific data. A total of 11405 known drug–target relationships were extracted from the database DrugBank, and five cancers and their disease-causing gene data were obtained from the human Mendelian genetic database. Through the direct neighbor method and using the tissue-specific interaction network as the background network, five tissue-specific drug–disease bipartite networks were constructed, which provided potential drug–disease associations. The results were verified by the CTD (comparative toxicogenomics database) standard. The experimental results show that the accuracy of drug repositioning studies based on tissue specificity and direct neighbor measurement will provide a reliable candidate set for in vivo and in vitro experiments of new drugs, which also provides new ideas for studying drug repositioning.

Keywords drug repositioning, tissue specificity, drug targets, disease genes, direct neighborhood measurement

Liang YU was born in 1979. She received her Ph.D. degree in Computer Architecture from Xidian University, Xi'an, China, in 2011. Currently, she is an Associate Professor at Xidian University. Her research interests include computational bioinformatics, graph mining, and network medicine.

Jin ZHAO was born in 1993. He received his M.S. degree in Computer Application from Xidian University, Xi'an, China, in 2017. His research interest focuses on network medicine.