基于复杂网络的合成致死预测方法研究综述

刘 闯[1),2)] 舒胜利[1)] 詹秀秀[1),2)] 张子柯[3)]

[1)](杭州师范大学阿里巴巴复杂科学研究中心 杭州 311121)

[2)](移动健康管理系统教育部工程研究中心 杭州 311121)

[3)](浙江大学传媒与国际文化学院 杭州 310028)

摘 要 合成致死(Synthetic Lethality,SL)是一种负遗传相互作用,描述的是两个非必要基因之间的相互关系:其中任何一个基因的突变对细胞存活的影响很小,但两个基因的共同突变会导致细胞死亡或其他有碍细胞存活的表型. SL 对于解释复杂生物过程、推动癌症的临床诊治有着重要的意义. 因此,利用海量的高通量数据,通过构建数据分析模型和计算方法,从计算的角度进行 SL 对的挖掘和预测,是计算生物学研究的一个重要方向. 本文首先对于 SL 预测所使用的相关数据进行了详细的综述,然后从生物网络这一全新视角出发,重点讨论了基于网络分析的 SL 预测方法. 从网络上的统计学方法、基于网络结构变化的方法、基于网络特征学习的方法、基于图表示学习的方法四个方面综述了相关预测模型和研究的最新进展,详细地比较了各类方法的算法思路、应用场景和优缺点,最后针对 SL 预测的结果评估和验证方法的研究进展进行了论述. 在此基础上,论文进一步总结出 SL 预测研究中所面临的几项挑战,并针对性的对未来发展方向进行展望,希望为今后的相关研究提供一些有用的参考和思路.

关键词 合成致死;复杂网络;基因突变;机器学习;预测方法

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2023.01670

Review of Network-Based Methods for Synthetic Lethality Prediction

LIU Chuang[1),2)] SHU Sheng-Li[1)] ZHAN Xiu-Xiu[1),2)] ZHANG Zi-Ke[3)]

[1)](*Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou* 311121)

[2)](*Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou* 311121)

[3)](*College of Media and International Culture, Zhejiang University, Hangzhou* 310028)

Abstract Tumors usually arise from loss-of-function mutations or gain-of-function mutations in oncogenes, which are usually found only in cancer cells, and it has been of great interest to exploit this weakness of cancer cells to develop more effective means of fighting cancer. Synthetic lethality (SL) has been proposed in this context for cancer treatment. Synthetic lethal is a negative genetic interaction that describes the interrelationship between two non-essential genes: the loss of either gene has little effect on cell survival, but the joint loss of both genes leads to cell death or other phenotypes that are poor for cell survival. Synthetic lethal interactions have an important role in explaining complex biological processes and the treatment of human cancers. In order to identify more synthetic lethal pairs and make the concept of synthetic lethality benefit the cancer population, researchers have gone through a process from SL screening in model organisms to SL screening in human cells. However, the high cost of experimental screening and the frequent off-target problems make in vitro screening of synthetic lethal difficult. Therefore, how to mine and predict synthetic lethal pairs through massive high-throughput data analysis is an important direction of

收稿日期:2022-07-01;在线发布日期:2023-01-05. 本课题得到国家自然科学基金项目(61873080,92146001)、国家社会科学基金重大项目(19ZDA324)、中央高校基本科研业务费资助. **刘 闯**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为复杂网络、生物信息. E-mail: liuchuang@hznu.edu.cn. **舒胜利**,硕士研究生,主要研究方向为复杂网络、生物信息. **詹秀秀**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为机器学习、复杂网络. **张子柯**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为智能计算、计算传播. E-mail: zkz@zju.edu.cn.

synthetic lethal-related research in recent years. In this paper, we focus on synthetic lethality prediction based on network analysis and review recent advances in relevant prediction methods and models in four areas: statistical methods on networks, methods based on the variation of network structure, methods based on network feature learning, and methods based on graph representation learning. We compare in detail the computational ideas, application scenarios, advantages and disadvantages of various methods, and analyze and summarize the main challenges and possible directions of development for synthetic lethality prediction. The statistical methods on networks are transformed from experimental screening, and the biological characteristics of SL distinguished from non-SL pairs are presented in the form of statistical judgments as the screening conditions to complete the prediction task; the prediction methods based on network structure changes take the main idea of simulating the occurrence of synthetic lethal in the network and quantifying its impact on the whole network; the methods based on network feature learning use traditional machine learning methods in the prediction task, and the methods based on graph representation learning, which would be the main direction of future research, represent the network nodes as an low-dimensional vector for the synthetic lethal prediction task. In general, biological network-based synthetic lethal prediction methods are still in the developmental stage, especially for human cells. And the synthetic lethal research faces many challenges: first, SL data consists of a small number of positive samples and a large number of unlabeled samples, and the extremely uneven nature of the data is not a small challenge for all types of computational methods; second, starting from the biological background of SL itself, the same pair of SL presents different phenotypes in different cancers, and this specificity makes its clinical use limited, so it needs special attention from researchers. This specificity makes its clinical use limited and therefore requires special attention from researchers. To address the above challenges, we propose several possible research directions, such as finding reliable negative samples from unlabeled samples for prediction tasks by PU learning heuristics; modeling SL data by knowledge graphs and graph neural networks to use small amount of reliable SL data and other auxiliary biological information as efficiently as possible; minimizing synthetic lethal specificity in research by using single-cell data, etc. This paper may offer some ideas and guidelines for the research of synthetic lethal, and we hope that it will draw increasing interdisciplinary attention from computer scientists, biologists, physicists and so on.

Keywords　synthetic lethality; complex networks; genetic mutation; machine learning; prediction methods

1　引　言

全球癌症发病数和死亡数持续增长，癌症防治形势面临严峻挑战，加紧研发有效的治疗手段变得刻不容缓[1]. 目前，最常用的癌症临床救治方法是向患者输送高剂量的、有辐射的化学物质杀死体内癌细胞. 但这种治疗手段对人的体细胞毫无选择性，在杀死癌细胞的同时也会对正常组织细胞造成严重损害，引发脱发、多器官衰竭等并发症[2].

肿瘤是细胞在发育过程中由于基因突变而引发其无限增殖的一种病变，其突变主要分为两种：致癌基因的功能获得性突变和抑癌基因的功能丧失性突变[3]. 这些突变只发生在癌症细胞中，是癌细胞的弱点，如何利用这一弱点找出更有效且安全的治疗手段是一个值得研究的问题[4]. “合成致死”(Synthetic Lethality，SL)作为一种新的抗癌策略引起人们的注意. “合成致死”[5-6]指两个非必要基因双突变(会造成基因的低表达或过表达)造成细胞死亡，而它们分别突变却不会对细胞生长有影响的生物学现象. 随着研究的深入，在该概念上延伸出合成剂量致死(Synthetic Dosage Lethality，SDL)的概念[7]，SDL指一个基因突变导致基因过表达使得另一个基因成为必需的现象，如图 1 所示.

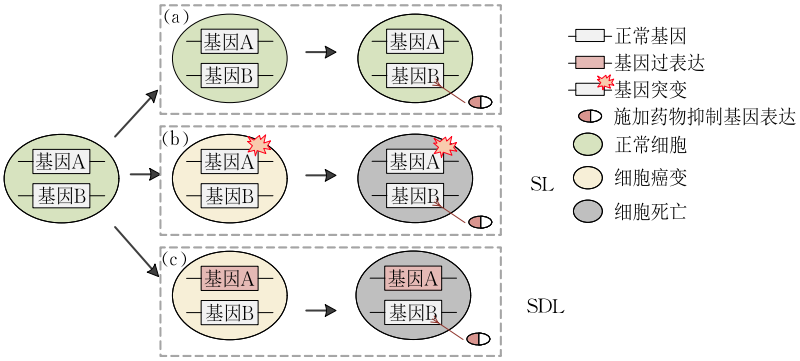

图 1 SL 及 SDL 概念示意图((a)正常细胞不存在肿瘤特异性突变,所以当抑制其中一个基因表达时,不会影响细胞的正常存活状态;(b)SL 是指两个非必要基因双突变造成细胞死亡而它们分别突变却不会对细胞生长有影响的生物学现象;(c)SDL 指的是一个基因突变导致基因过表达使得另一个基因成为必需的现象.因为(b)SL 和(c)SDL 现象的存在,如果一个基因与一个肿瘤特异性突变的基因具有 SL 或 SDL 作用,那么该基因将是一个潜在的抗癌药物靶点)

基于 SL 的精准靶向治疗是癌症治疗领域的一个重要突破,为癌症的新型治疗带来了曙光[8],如果一个基因与一个肿瘤特异性突变的基因具有 SL 作用,那么该基因将是一个潜在的抗癌药物靶点[7].如 BRCA1/2 和 PARP 作为一对典型的 SL 对可以治疗乳腺癌,它们是细胞内修复脱氧核糖核酸(DeoxyriboNucleic Acid,DNA)损伤的两种不同路径,正常情况下细胞内两条修复路径都可起作用,但对于 BRCA1/2 基因失活的肿瘤细胞而言,只能通过 PARP 修复 DNA 的损伤.在使用 PARP 抑制剂后,PARP 修复途径被阻断,肿瘤细胞因 DNA 损伤而死亡,但是正常细胞由于有另一条备用修复途径而不受影响[9](如图 1 所示).

SL 与癌症表型之间具有关联关系,存在 SL 关系的基因对同时突变会定向性地杀死细胞,“SL 用于癌症治疗”[10]的观点被提出来.为了发现更多的 SL 对,人们经历了从模式生物中的实验筛选鉴定[11-13]到人类细胞中的实验筛选鉴定[14-16]的过程.一方面,生物实验面临着脱靶现象频发的问题;另一方面,全基因组规模的筛选鉴定是一项高成本的任务,如人类的基因数目大约为 25 000 个,产生的基因对数目可达 3 亿对.在这样的背景下,人们开始将视线转移到由生物实验积攒的大量数据上来,希望从这些高通量数据上获得更多潜在人类 SL 对的线索,在深入探索 SL 相互作用机制的基础上,为临床治疗提供更多可能性.但是,高通量数据有着数据规模庞大、类型复杂的特点,“使用什么数据”和“怎么使用数据”是高通量数据用于 SL 预测的两个重要问题.

一种可能的尝试是基于进化保守性的同源映射[17-18].模式生物的全基因组基因数量较少,对于其上的 SL 筛选鉴定任务已经基本完成.人和模式生物来源于共同的祖先,进化过程中会保留相当一部分相似的基因序列,通过比对基因测序数据可以把模式生物中的 SL 对映射到人类细胞中,从而发现潜在 SL 对.但是,后来的研究发现,人类与酵母(最常用的模式生物)间的同源基因的比例不到 50%,基于模式生物的进化保守性推测人类 SL 对的想法[18-19]面临挑战.

另一种尝试是以基因功能分析为导向的 SL 预测研究[20-21].从细胞代谢角度或信号通路角度出发,分析基因的进化特点或与细胞生存之间的关系,从而推测基因在癌变过程中可能发挥的作用,以此发现可能存在基因功能互补的基因组合.但是,SL 作用是一种复杂的生物反应,其背后涉及包括基因、蛋白质、代谢物等多种分子间的相互作用关系,仅从基因功能角度分析是片面的,不利于 SL 的挖掘和进一步认识.

网络是自然界中大多数复杂系统在逻辑上的存在形式,网络中的节点表示系统中的各个组件,节点间的连线描述组件之间的相互作用关系.网络作为一种数据组织形式,可以描述不同生物实体(如基因、药物、疾病等)间的多种抽象关系(如调控、相互作用、功能相关等)[22],是大多数生物数据的组织形式.当整合多个分子类型的生物数据研究 SL 关系时,一方面可以尽可能还原 SL 基因在体内所处的生物微环境,有助于深入且系统的认识 SL 关系;另一方面可以借助丰富的数据信息获得高效的预测精度.当专注于研究单一分子网络时,可以借助网络科学和数据挖掘等方法将 SL 关系投射到生物网络中,进行由局部到整体的分析,方便揭示 SL 基因对在该网络中的特殊拓扑模式,还为研究者提供了由微观到宏观的视角转变,以一种非常全面的方式认

识和理解 SL 关系，对于解释其机理是非常重要的. 另外，基于网络模型的 SL 研究具有相当大的灵活性，当人为改变网络中节点所代表的生物实体或分子时，则可以为研究者提供不同粒度的关于 SL 的认识和见解. 最后，相比于湿实验室鉴定来说，基于网络的 SL 预测研究的发现过程有着更高的效率，最大化生物数据中潜在价值的同时，将人力和物力等资源的损耗降到了最小.

因此，本文将重点关注基于网络模型的 SL 预测方法，探索生物网络分析在 SL 预测上的发展方向与应用前景. 文章的组织形式如下，首先简单介绍 SL 预测中的常用数据；其次，重点综述不同种类下的 SL 预测方法；接着，概述 SL 预测结果的评估和验证方法；最后，论述 SL 预测研究面临的挑战和未来的发展前景.

2　数据描述

丰富的数据是 SL 研究的基础，不同的方法对于数据的需求不同. 在 SL 预测中，常用的数据有两类：一类是 SL 标签数据，即通过实验确认或是生物推断的方法得到的已知 SL 对；第二类是特征数据，即基因之间存在 SL 关联的生物信息，本文主要介绍基因组学数据、转录组学数据、蛋白质相互作用数据和基因功能注释数据.

2.1　SL 标签数据

对于 SL 预测任务来说，我们更加关注存在 SL 关系的基因对，因为足够数量的可信正样本对于监督学习算法的构建至关重要. Syn-Lethality[23] 是第一个收集人类癌症相关 SL 对的综合性数据库，它将通过实验发现和验证的人类 SL 对以及酵母 SL 对的人类同源基因对整合到一个网络中，并与基因功能、细胞通路和分子机制相关的注释相关联. 在这之后，SynLethDB[24] 以 Syn-Lethality 为基础，从多个来源收集了包括人、酵母、果蝇等生物在内的 SL 对. SynLethDB 不仅拥有同 Syn-Lethality 一样丰富的多组学数据，而且富含多个物种的互作信息，最近更新的版本中，还增加了 SynLethKG 知识图（KG4SL[25] 方法中使用）. 因此，SynLethDB 已经成为目前研究中使用最多的开源数据库，多项基于该数据库的研究结果表明这种选择是合理且有效的. 另外，还有一些其他研究整理的 SL 数据也常被使用，如表 1 所示（表 1 中只列出明确包括正负样本的数据集）.

表 1　SL 标签数据集

数据库	描述	SL 基本信息	时间
Syn-Lethality[23]	第一个人类 SL 综合数据库	关于人类泛癌背景下的 2529 对 SL	2014
Jerby-Arnon 等人[26] 的研究	作者预测算法得到的人类 SL	关于人类泛癌背景下的 2816 对 SL 和 3635 对 SDL	2014
Lee 等人[27] 的研究	多个研究中大规模筛选结果	关于人类泛癌背景下的 154707 条 SL	2015
Lu 等人[28] 的研究	其他研究的实验测量结果	关于人类的 270 对 SL 和 5660 对非 SL	2015
Shen 等人[29] 的研究	CRISPR-Cas9 组合筛选研究	关于人类不同细胞系下的 152 对 SL（293T：59 对；A549：57 对；Hela：52 对）	2017
Slorth[30]	作者预测算法得到的人类 SL	关于 5 个物种的 207 921 对 SL（人：51 863 对；酿酒酵母：37 256 对；黑腹酵母：9300 对；世宗酵母：52 594 对；秀丽隐杆线虫：56 908 对）	2019
SynLethDB-v2[24]	人类 SL 的综合性数据库	关于 5 个物种的 50 868 对 SL（人：35 943 对；酿酒酵母：14 000 对；小鼠：381 对、黑腹果蝇：439 对；秀丽隐杆线虫：105 对）	2020

SL 预测模型通常是监督学习，需要正负样本共同完成模型的训练，负样本的准备是该预测任务上的一个难题. 最简单的做法是从未知样本中随机抽取和正样本同等数量的样本作为负样本，这种做法很常用[30-32]，且可以获得足够数量的负样本. 但是，随机抽样的负样本可能会引入假阴性，另一种被采用的做法是根据基因相互作用（Gene Interaction，GI）得分选择正负样本，如 Hao 等人[33] 的研究中选择 GI 得分小于 −3 的基因对为正样本，而提取 GI 得分接近 0 的基因对为负样本；而 Wan 等人[34] 的研究则依据 GEMINI[35] 算法得分来选择负样本.

2.2　特征数据

SL 现象对于癌症治疗的现实意义和潜在价值使得“SL 对的预测”成为生物信息学研究上的一个热门研究课题；同时，SL 作为一种复杂的生物学现象，人们至今无法清楚的明白其完整的形成和作用链条. 故对于 SL 基因对的研究常常需要整合来源于高通量测序的、刻画不同生物分子性质和功能的不同类型的生物数据. 在这里，我们认为常用于 SL 预测研究中的特征数据包括基因组学数据、转录组学数据、蛋白质相互作用数据和功能注释数据，相关数据库见表 2.

表 2 特征数据集

类别	常用特征	数据库	描述	最近更新
基因组学	突变覆盖率[36]、杂合/纯合缺失发生率[28]、基因序列相似性[37-38]	Genbank	包含 2 亿多条信息的 DNA 序列数据库	—
基因组学、转录组学	—	TCGA	包含 33 种癌症相关的基因组、转录组、表观遗传、蛋白组等各个组学数据和患者临床信息的综合性数据库	2022.3.3
转录组学	差异表达[39]、共表达[39]、特殊表达模式[28]	GTEx	包含 42 种不同组织类型下的人类正常组织的基因表达信息的数据库	2022.5.12
		LINCS L1000	存储人的细胞系在小分子干扰、基因敲除等扰动下的基因表达情况的数据库	—
蛋白质相互作用数据	最短路径长度[30]、共享的n阶邻居数目[30]、PPI 拓扑相似度[33,40-41]	STRING	包含 14 094 个物种的 67 592 464 个蛋白质之间的超过 200 亿条 PPI 的数据库	2021.8.12
		BioGRID	来自多物种的 2 496 557 条与基因、蛋白质和化合物相关的相互作用的数据库	2022.5.9
		HIPPIE	包含 16 835 个蛋白质相关的 287 357 条高可信度的人类 PPI 的数据库	2022.4.29
		HPRD	包含人类蛋白质表达谱、结构域、亚细胞定位、转录后修饰、通路、相互作用等信息的综合性数据库	—
基因功能注释	共享的通路/基因功能数目[37-38]、GO 语义相似度[33,40-42]、功能相似度[39]	GO	从生物过程、细胞成分和分子功能三方面描述特定基因的功能的数据库	2022.5.16
		KEGG PATHWAY	包含分子相互作用、反应和关系网络的通路图集合	2022.4.1
		CTD	整合大量化学物质、基因、功能表型和疾病之间相互作用的数据库	2022.5.4
		Reactome	描述不同实体(基因、蛋白质、复合物、疫苗、抗癌疗法和小分子)参与生物反应的通路数据库	2022.4.5
		MsigDB	从位置、功能、代谢途径、靶标结合等角度出发,为人类基因构建不同的基因集合的数据库	—

2.2.1 基因组学数据

基因是控制生物性状的基本遗传单位,基因组学是对生物体全基因组的结构、功能、进化等的研究.对于基因组数据的挖掘可以帮助理解 DNA 分子及其产物之间的功能关联,有助于提高对包括 SL 在内的、产生重要表型的生物学现象的认识[43].

由 SL 定义可知,基因突变是发现 SL 关系的一个重要因素,Ye 等人发现 SL 对间发生突变的概率更高[36].除此之外,研究显示基因组结构变异与 SL 基因对之间的特殊作用有关[7],其中,体细胞拷贝数变异(Somatic Copy Number Variation,SCNV)经常被用到 SL 预测任务中.如在网络上的统计学方法中,通常将观察到的 SCNA 作为一个重要的判断依据[26,44].在基于网络特征学习的方法中,除了利用基因突变数据刻画 SL 基因组合之间的关系[28,45]外,还会使用基因序列数据,寻找两个基因之间的相似性[37-38].值得一提的是,癌症基因组图谱(The Cancer Genome Atlas,TCGA)数据库[46]因其富含基因组学、转录组学、蛋白组学、表观遗传学等各种组学数据,成为了癌症研究中使用最频繁的数据库之一.

2.2.2 转录组学数据

转录组是指一个活细胞可以转录的所有核糖核酸(Ribonucleic Acid,RNA)的总和,基因表达的产物是所有功能性 RNA,是转录组的重要组成部分,故转录组学是 RNA 水平的基因表达研究.转录组图谱提供什么基因在什么条件下表达的信息,经常搭配基因组学数据一起使用,为的是尽力还原 SL 基因所处的肿瘤微环境,这对于提升 SL 预测任务的性能、推断 SL 基因的功能、揭示 SL 关系的作用机制至关重要.

SL 的发生与基因的异常表达紧密相连,结合转录组数据设计和开发预测模型可以提供更多的生物背景参考信息.基于统计学的方法大多需要转录组数据的支持,如 DAISY 方法考虑了基因表达与基因突变的关系,基于"一个突变基因的 SL 伴侣在携带这种突变基因的原发性肿瘤样本中表达水平会更高"的假设发现新的 SL[26].EXP2SL 则借助 LINCS L1000[47]中不同细胞系的基因表达谱学习基因在特定遗传背景下的相互作用关系用于 SL 预测[34].最近的一种新尝试是从正常人类细胞的基因表达数据中预测新 SL 互作[48].

2.2.3 蛋白质相互作用数据

蛋白质是生命的物质基础,与各种生命活动密切相关.通过探索蛋白质群体的作用方式、功能机制、调控水平,可以更全面地了解复杂生物学现象的调节过程,揭示生命活动的基本规律.

在 SL 预测中最常使用的蛋白质组学数据为蛋白质相互作用(Protein-Protein Interaction，PPI)网络. 因为基因编码蛋白质的映射关系存在，PPI 网络有时可以作为基因相互作用网络的补充. 如 Kranthi 等人[49]曾将来自 HPRD 数据库[50]的 PPI 网络映射为基因相互作用网络，并观察网络扰动前后网络效率的变化来识别新的 SL 对，类似的处理办法也可以在 Magen 等人[51]的研究中看到. 另外，蛋白质在一定程度上维持着许多关键的生物过程，同时也反映了许多重要的生物功能. 为了有效利用 PPI 上的丰富信息，更多的研究选择 PPI 网络作为描述基因关系的功能网络. 如 Paladugu 等人[52]和 Benstead-Hume 等人[30]都是从 PPI 网络上提取了一系列图论特征来刻画正负样本之间的潜在关联；Chua 等人[53]提出的基于 PPI 网络的 FSWeight 也被多个研究[33,40-41]所采用；基于图卷积神经网络和图自编码器的多个算法中同样出现了 PPI 网络的身影[33,41,54].

在 SL 预测研究中，PPI 有着举足轻重的地位，很多研究者会选择将 PPI 作为整合其他数据的媒介，与 PPI 相比，其他类型的生物数据通常作为补充数据存在. 关于 PPI 网络的公开数据资源丰富，如 STRING[55]包含了 5090 个物种、24 584 628 个蛋白质与 3 123 056 667 种相互作用关系，BioGRID[56]收集了来自模式生物和人类的蛋白质互作关系. 表 2 中列举了 SL 预测上使用较多的几个数据库.

2.2.4　功能注释数据

功能注释是一种分析技术，是将生物功能信息附加到基因或蛋白质序列的过程，借助功能注释数据可以认识生物分子在生物体中的功能，刻画生物分子在生物系统中所处的位置，并对其进行相应的分类.

过去的研究表明，SL 对可能会更容易参与相似或相同的生物功能[7,10,19]. 故大多数研究都选择功能注释数据作为提高模型性能的补充数据. 如在基于网络特征学习的方法中，基因对共享的生物注释数量是经常出现的关系描述[30,38]；另一个总是被研究者们使用的指标为基因本体(Gene Ontology，GO)语义相似度[33,40-42]. 在图表示学习方法中，功能注释数据也会作为一张或多张单独的网络输入到模型中[25,33,41]，多个模型的预测结果表明添加功能注释后的模型精度更高.

最著名的功能注释数据库是 GO，它将基因的功能划分为三类术语：生物过程(Biological Process，BP)、分子功能(Molecular Functions，MF)和细胞成分(Cellular Components，CC)[57]. 更多的功能注释数据库见表 2.

3　基于网络的预测方法

基于网络模型的 SL 预测算法框架如图 2 所示，

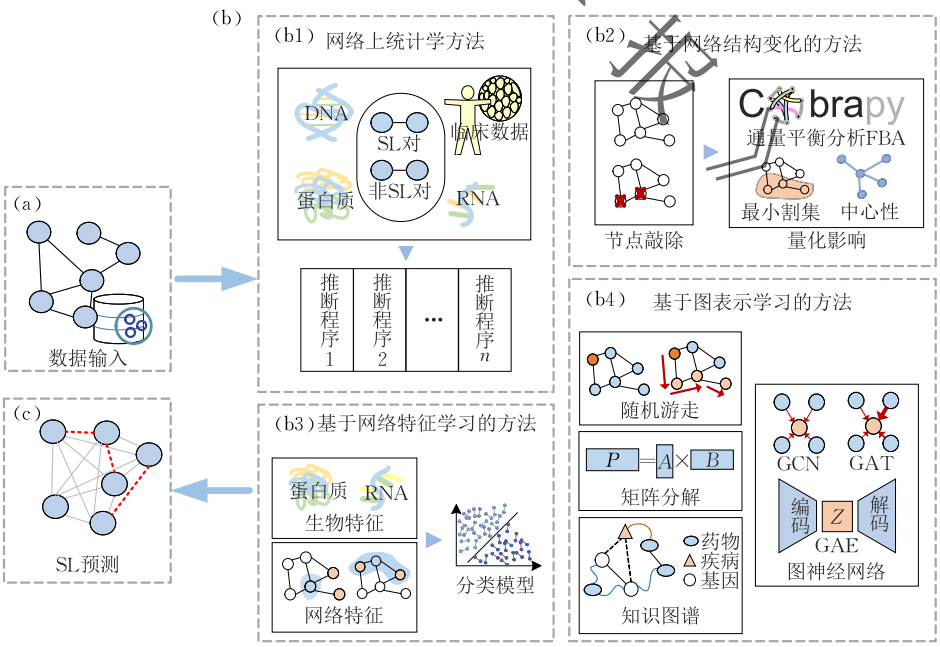

图 2　基于复杂网络的 SL 预测方法框架图((a)选择合适的数据集作为后续方法/模型的输入；(b)根据已有的数据集选择适当的框架来实现方法/模型；(c)基于构建的方法/模型完成 SL 预测任务. 其中，本文将基于复杂网络的 SL 预测方法分为四类：(b1)网络上的统计学方法、(b2)基于网络结构变化的方法、(b3)基于网络特征学习的方法和(b4)基于图表示学习的方法)

相关算法都是将 SL 预测看成一个二分类问题来进行处理,即预测两个基因是否为 SL 对.首先是选择合适的生物网络数据和组学数据作为算法模型的输入(图 2(a)),然后是构建基于网络的预测方法和模型(图 2(b)),最后是得出潜在的 SL 预测结果(图 2(c)).根据方法特点将其分为 4 类:网络上的统计学方法、基于网络结构变化的方法、基于网络特征学习的方法和基于图表示学习的方法.前两种方法整合多种特征数据的方式是依据生物观察或经验结果设计具体的判断条件或指标,基于网络特征学习的方法则是以特征工程的方式完成多类数据的整合,基于图表示学习的方法选择将多个分子水平的生物数据作为 SL/PPI 网络中节点的属性值参与网络分析和模型构建的过程.

需要强调的是,基于网络模型的 SL 预测指的是通过已知的生物网络节点以及网络结构等信息,完成两个基因是否是合成致死关系的二分类判断任务,任务本身强调已知关系的划分,并非链路预测问题中的未知关系的补全(有少数研究将 SL 预测研究的输入输出都置为 SL 网络,完成 SL 网络上的链路预测,例如基于矩阵分解的 SL 预测研究[40]).通常来说,大多数研究者会选择让 PPI 网络作为承载其他类型生物数据和发现潜在拓扑关系的媒介,如图 2 中(a)图所示.在预测阶段,发现新 SL 关系的过程是建立在相关基因对的预测中,如图 2 中(c)图所示.另外,SL 预测研究中一个值得注意的问题是,SL 网络和 PPI 等生物网络之间并不存在直接的关联关系,即存在 SL 关系并不表示存在基因相互作用关系,存在基因相互作用关系的并非存在 SL 关系.

3.1 网络上的统计学方法

网络上的统计学方法的基本思想是将 SL 对区别于非 SL 对的生物学特点以统计学判断的形式呈现为筛选条件(如图 2(b1)所示).参与筛选的基因组合以串行或并行的形式依次通过每一轮测试,最终选择通过所有测试的基因组合为最后的预测结果.即:

串行:

$$f_{i,j\in V,i\neq j}(i,j)=f^1_{i,j\in V,i\neq j}(V,i,j)\rightarrow\cdots\rightarrow f^n_{i,j\in V^{n-1},i\neq j}(V^{n-1},i,j) \quad (1)$$

或并行:

$$f_{i,j\in V,i\neq j}(i,j)=f^1_{i,j\in V,i\neq j}(V,i,j)\cap\cdots\cap f^n_{i,j\in V,i\neq j}(V,i,j) \quad (2)$$

其中,V 通常为待测试基因网络中的不重复基因列表,$f_{i,j\in V,i\neq j}(i,j)$ 表示判断基因组合(i,j)是否为 SL 需要经历的选择过程.

对于串行设计的方法而言,筛选通常表现为有顺序的逐轮筛选,第 n 轮筛选 $f^n_{i,j\in V^{n-1},i\neq j}(V^{n-1},i,j)$ 的输入 V^{n-1} 为上一轮筛选的输出,最终筛选的结果为依次通过每一轮筛选的基因组合;对于并行设计的方法而言,筛选通常表现为无顺序的集体筛选,每一轮的筛选输入都为 V,最终的筛选结果为通过所有轮次筛选的并集.

由上可知,网络上的统计学方法的关键是根据 SL 概念的生物学观察或已有的经验结果,基于不同分子层面的数据设定合适的统计筛选条件.在已知的该类方法中,常常用于设计选择过程的依据包括两点:(1)SL 对更有可能参与相同或相似的基因功能或生物通路,表现出功能备用的关系[7,10,12];(2)在没有发生 SL 的原发性肿瘤细胞中,一个基因有突变或表达不足时,会引起其 SL 伙伴的重要性增强,且这种重要性可以通过该伙伴基因的表达水平和 SCNA 来观察.具体来说,在基因 i 发生突变或表达不足的原发性肿瘤样本中,i 基因的 SL 伙伴 j 基因的表达会更加活跃或 j 基因更容易扩增.

科研人员基于以上认识提出了多种 SL 预测方法[26-27,44].其中,最经典的方法为 DAISY[26],它结合基因表达、基因突变、基因重要性筛选结果得到了含 2077 个基因的 SL 相互作用网络.对于一对基因(i,j),必须按顺序通过以下三个串行判断才可以被判定为 SL 对:在基因 i 不活跃的样本中,基因 j 的 SCNA 显著高于基因 i 正常表达的样本;在基因 i 不活跃的样本中,基因 j 的基因重要性显著高于基因 i 正常表达的样本;基因 i 和基因 j 的系统发育相似性(基因功能上的认识)呈现出显著的正相关.当将前两个条件中的基因 i 的表达状态修改为过表达时,同时通过以上三个条件的基因对即为 SDL.DAISY 方法的提出是借助统计推断预测 SL 的初次尝试,是 SL 预测工作上的重要进展,证明了基于生物先验知识设计统计筛选完成预测任务的可行性.

受 DAISY 的启发,MiSL 方法[44]被提出来.与 DAISY 不同的是,MiSL 选择使用更加宽松的"并行+串行"的筛选策略,同时将观察视角从 DAISY 的基因表达状态上转移至基因表达的变化趋势上.对于基因对(i,j),在 i 基因突变导致的异常表达的样本中,观察 j 基因的 SCNA,如果 j 基因在肿瘤样本中有更高频率的扩增或更低频率的删除,并且基因 j 的表达变化趋势和扩增或删除的趋势一致,则认为这对基因更有可能存在 SL 关系.MiSL 方法克服了 DAISY 方法利用少量失活突变的保守性,扩大了潜在 SL 的范围.且 MiSL 抛弃了基因重要性筛

选数据的使用，杜绝了其上的假阴性[58]问题. 同时 MiSL 可以针对特定癌症类型的特定突变进行分析，是相对于 DAISY 的改进和拓展.

值得一提的是，Lee 等人[27]为了进一步找出对于患者的治疗有着更大积极作用的 SL 对(cinical Synthetic Lethality，cSL)，提出 ISLE 方法. 同 DAISY 一样，该方法串行地执行三个推断规则，同时通过三轮筛选的基因对为目标基因对，但 ISLE 的输入是实验验证或者计算推断的 SL 组合，该方法更加强调从 SL 中找出 cSL，并非发现新的 SL 组合. 另外，最近的研究开始关注特定癌症类型上的 SL 对的探索，如 ASTER[59]继承 DAISY 的方法设计，认为“一个基因表达不足时另一个基因的重要性则上升”，但二者的不同点是对于表达不足的衡量，ASTER 方法认为肿瘤细胞中的高低表达应该来源于非癌细胞表达数据的对比，故 ASTER 有意避免了突变数据的使用，减少了数据中的噪音，ASTER 找出了存在于胃癌和乳腺癌中的 SL 对. Yang 等人[39]的 SiLi 方法则从肝癌患者的基因表达和突变等信息中找出了肝癌特异性的 SL 对，该方法分别考虑了基因对间功能相似度关系、基因对间差异表达关系、基因对间共表达关系、基因对在生存分析上的表现及聚合前四个关系的聚合相关性，相比于其他方法，SiLi 考虑了更多的可能性，方法设计更加开放，但通过的 SL 对存在假阳性的可能性更高.

总之，基于统计学的方法作为一种无监督方法，避开了特征提取和特征选择，计算复杂度较低；同时也没有训练过程，避免了负样本采样困难的问题. 另外，基于统计学的方法与生物先验知识充分关联，故方法的可解释性较强. 但是，该类方法通常需要组学数据的支持，数据类型的多样性带来了数据处理上较高的操作复杂度，也更易受组学数据中的噪声干扰. 更加重要的是，该类方法要求研究者对 SL 的表达特点都有着妥当的生物学理解，以此保证筛选条件的合理性和正确性. 一般来说，这类方法对输入数据敏感，一方面需要足够多的样本数据支持，另一方面筛选条件依托于输入数据而存在，当改变数据源时，必须依次修改统计测试，且修改后的方法性能往往会变差，方法鲁棒性低. 关于文中提到的所有网络上的统计学方法见表 3.

表 3　网络上的统计学方法

方法	研究任务	算法描述	使用数据
DAISY[26]	人类 SL(泛癌)	基于三个串行的统计判断筛选泛癌背景下的人类 SL，其预测结果收录于 SynLethDB[24]中；建立了网络上的统计学方法的范式，为后续研究提供诸多参考的经典方法；但无法避免基因重要性筛选结果带来的假阴性问题.	SCNA、基因表达、基因突变、基因重要性筛选结果
MiSL[44]	人类 SL(泛癌)	基于串行＋并行的统计判断筛选泛癌背景下的人类 SL，方法“以突变为中心”，其重点是识别存在特定基因突变的基因的 SL 伙伴；方法克服了 DAISY 中基因因为突变不充分表现造成的假阴性问题.	基因突变、SCNA、基因表达
ISLE[27]	人类 cSL(泛癌)	基于显著共失活、更好的患者预后、基因对共同进化的假设设计的三个串行筛选过程，发现泛癌背景下的、更可能对临床治疗有作用的人类 SL. 方法首次关注 SL 在应用上的问题，其设计假设成为验证预测 SL 的标准.	SCNA、基因表达、临床数据、基因发育谱.
ASTER[59]	人类 SL(胃癌、乳腺癌)	基于患病和无病样本间基因表达的关系设计的三个串行统计判断筛选胃癌和乳腺癌背景下的人类 SL；有意避免突变减少了因为癌症样本间的差异带来的假阴性问题.	SCNA 和基因表达、非癌症样本的基因表达水平
SiLi[39]	人类 SL(肝癌)	基于功能相似度分析、差异基因表达分析、基因两两共表达分析、两两生存分析和等级聚合分析设计的寻找肝癌背景下人类 SL 的方法；补充了肝癌相关的 SL 的列表.	基因表达、基因突变、患者生存和 GO

3.2　基于网络结构变化的方法

SL 现象的发生对细胞的生长会产生不利甚至致命效果，从网络的视角看，该作用会表现为对整个网络系统的负影响，故可以通过将潜在 SL 对的两个基因节点从基因网络中去除，并根据该网络结构变化对于整个网络的影响来判断两个基因间的关系(如图 2(b2)所示). 去除节点后，该节点所表示的基因表达减弱或消失，可以模拟基因突变导致的不正常表达. 即当一对节点从网络中敲除后，SL 现象在该网络系统内发生，此时预测未知 SL 的关键问题是“如何正确反映并量化该作用给整个网络带来的不利影响”.

3.2.1　中心性度量

Kranthi 等人[49]曾提出“图信息中心性”的概念来描述整个网络系统在去除节点对后的影响. 假设网络 G 中的信息沿着最短路径进行传播，那么可以使用 $E(G)$来描述网络效率：

$$E(G)=\frac{1}{N(N-1)}\sum_{i\neq j\in G}\frac{1}{d_{ij}} \tag{3}$$

其中 N 为网络节点总数，d_{ij} 为节点 i 到 j 的最短距离. 敲除节点对(i,j)后的网络为 G'，敲除后网络效率的下降可以使用图信息中心性 C_{ij} 描述：

$$C_{ij}=\frac{\Delta E}{E}=\frac{E(G)-E(G')}{E(G)} \tag{4}$$

将该概念用于人类 PPI 网络上，在把所有节点对的中心性得分排序后，选择排名靠前的、对网络系统功能产生显著影响的基因对为潜在 SL 对.

3.2.2 通量平衡分析

通量平衡分析(Flux Balance Analysis，FBA)是一种广泛应用于生物化学网络上的数学模型，运用在代谢网络上可以计算代谢物的流量，从而预测一个有机体的生长速率或重要代谢物的生产速率. FBA 将代谢网络中浓度变化表示为化学计量系数矩阵 $\boldsymbol{S}$ 和参与代谢反应的酶通量 $\boldsymbol{v}$ 间的点积：$\boldsymbol{Sv}=\boldsymbol{0}$，右侧的零向量表示系统处于稳态；接着，需要根据系统的预测目标定义目标函数 $Z=\boldsymbol{cv}$，其中 $\boldsymbol{c}$ 表示参与反应的酶对整个预测的贡献；最后，使用线性规划来计算对应于稳态的通量解[60].

利用 FBA 可以寻找癌症的潜在药物靶点，因为通过自定义目标函数，FBA 可以计算扰动前后的细胞活力，以此来量化 SL 作用对网络稳健性的影响. IDLE 方法[61]利用 FBA 定量测量酶在人类代谢网络中的 SDL 效应，该方法是基于以下的观察：在一个代谢网络中，对于任意的 SDL 对(i,j)，当编码代谢酶 i 的基因表达降低时，编码代谢酶 j 的基因高表达会反过来抑制整个代谢网络系统的运作. 故可以通过酶 i 的敲除导致的酶通量减少模拟编码代谢酶 i 的基因低表达，并通过逐步增加酶 j 的酶通量来模拟编码代谢酶 j 的基因过表达，二者共同作用下的扰动与正常情况进行比较，使得生长明显减少的酶对的编码基因被预测为 SDL 对.

3.2.3 遗传最小割集

最小割集(Minimal Cut Sets，MCS)描述代谢网络中一组最小反应组或基因，这些反应组或基因的同时删除会消除网络执行特定任务的能力[62]. 当只在基因水平上考虑最小割集时，遗传最小割集(genetic Minimal Cut Sets，gMCS)的概念便应运而生. gMCS 指基因的最小子集，它的同时移除会直接阻断特定的代谢任务[63]. gMCS 的概念由 Apaolaza 等人[63]提出，并用来寻找高阶 SL 作用. 在癌症研究中，代谢反应可以理解为多个基因参与的生物通路过程. 由 gMCS 的概念可知，去除一组 gMCS 基因会让生物通路过程受阻，当受阻的生物通路过程与癌细胞的增殖相关时，gMCS 基因组的移除可以达到与 SL 作用相同的效果，中断癌症的进一步发展. Apaolaza 等人认为一组 gMCS 基因对应于一个高阶的 SL 作用，且当一个基因的 gMCS 伙伴都低表达时，这个基因会显示出相当大的重要性. 曾有研究表示 RRM1 是一个有希望的代谢靶点[64]，故 Apaolaza 等人[63]采用 Tobalina 等人扩展的 FBA 方法[62]配合基因表达数据，在人类代谢网络中寻找 RRM1 在多发性骨髓瘤细胞系中的高阶 SL 作用. 后来，Apaolaza 等人[65]发现肿瘤细胞的细胞外营养环境与 SL 的发生相关，于是将 gMCS 的概念扩展为营养-遗传最小割集(nutrient-genetic Minimal Cut Sets，ngMCSs)以识别导致细胞死亡的特定遗传环境和营养干扰相关的 SL 对. 这一次的尝试不仅为 MCS 赋予了新的实际意义，而且解释了体外鉴定体内 SL 时的抵抗性来源，对于 SL 预测和鉴定工作都是一项重要成就.

总之，基于网络结构变化的方法通过设计模拟实验观察并量化 SL 发生前后的影响力，根据影响力的强弱发现潜在的 SL 基因对. 在癌症相关的相互作用网络中去除某对或某些基因的作用类似于体外鉴定中基因沉默，故基于网络结构变化的方法更像是体外鉴定的计算机实现，对于解释“SL 对为什么成为 SL 对”更容易. 另外，得益于 MCS 概念的引入，人们首次尝试了高阶 SL 对的预测. 但是，基于网络结构变化的方法大多为代谢网络上的硅敲除，从基因编码产生酶到酶催化代谢反应，代谢网络与基因的关系是间接的，同时也是局部的，因为并不是所有的基因都会与参与代谢的酶相关. 关于文中提到的所有基于网络结构变化的方法见表 4.

表 4 基于网络结构变化的方法

方法	研究任务	算法描述	使用数据
Kranthi[49]	人类 SL(泛癌)	定义图信息中心性衡量基因对去除后对整个基因网络的影响来寻找泛癌背景下的人类 SL.	PPI、癌症基因列表
IDLE[61]	人类 SDL(泛癌)	在人类癌症代谢网络中改变由基因编码的酶对的酶通量后，使用通量平衡分析量化扰动前后对细胞活性的影响，由此发现泛癌背景下的人类 SDL.	人类代谢网络
gMCS[63]	人类 SL(泛癌)	根据 Tobalina 等人[62]的方法在代谢网络上发现 RRM1 相关的高阶 SL.	人类代谢网络、基因表达
ngMCS[65]	人类 SL(泛癌)	扩展 gMCS 来识别引发细胞死亡的特定环境遗传和营养干扰相关的 SL.	人类代谢网络

3.3　基于网络特征学习的方法

网络上的统计学方法利用生物网络知识和基因组数据，依据统计假设推断潜在 SL 对. 基于网络结构变化的方法模拟 SL 激活发现潜在 SL 对. 然而，它们都没有使用已知 SL 对信息挖掘潜在 SL 对. 基于网络特征学习的方法可以通过特征提取来完成多类生物网络数据间的有效整合，且整合过程十分灵活，不受生物数据类型的限制，任何在 SL 对和非 SL 对中表现出差异性的特点，经过量化之后都可以用于 SL 预测任务中. 该类方法将 SL 预测问题定义为二分类任务，通过构建监督学习分类器发现新的 SL 对(如图 2(b3)所示). 在这里，我们按照参与模型训练的特征类型将基于网络特征学习的方法分为三类：基于纯网络拓扑特征的方法、基于纯生物功能特征的方法和基于混合特征的方法.

3.3.1　纯网络拓扑特征

纯网络拓扑特征的提取将存在 SL 关系的两个基因抽象为网络中的节点，以复杂网络理论和方法为基础，分析两个基因在网络视角下的结构特点和相互关联.

在网络研究中，通过图论方法表征复杂网络的拓扑关系是研究网络中不同节点间关联的重要手段，这是因为网络的拓扑指标可以定性或定量地描述不同节点所代表的真实实体之间的相互作用关系. 如在基因网络中，一个基因的度描述为与其有直接相互作用的基因数. 某基因的度较大，则可以认为这个基因与其他基因之间的相互作用关系更加紧密. 基于以上认识，Paladugu 等人[52]借助度、聚类系数、紧密中心性、最短路径、介数中心性等来刻画酵母 SL 对在酵母 PPI 网络上的关联关系，采用支持向量机(Support Vector Machine，SVM)模型完成灵敏度和特异性均超过 85% 的预测任务. 类似的，Kranthi 等人[49]在人类 PPI 网络上重新定义“信息中心性”来推断人类 SL 对，用去除一对节点前后的网络效率下降与原网络效率的比值作为潜在 SL 对之间的关联程度，并且在前 100 的预测结果中发现了一部分被验证的 SL 对. Paladugu 等人的工作首次证明仅仅使用 PPI 网络上的图论特征就可以在一定程度上完成 SL 预测任务，这表明了 PPI 网络中的局部连通性和全局位置与基因功能特性相关. Kranthi 等人的工作进一步明确网络分析方法在理解 SL 作用上的重要性，这为后来 PPI 网络成为 SL 预测任务上最常用的数据提供了支持.

网络的拓扑结构会影响网络的功能，而网络的功能反过来又会影响网络结构的演化. 网络研究的另一个重要手段是结合复杂网络中的动力学理论刻画实体关系的关联性. 其中，网络扩散是最基本、应用最多的动力学过程，它由一连串随机的轨迹组成，可以表示一种不规则的变化形式. Qi 等人认为 SL 关系作为一种负遗传相互作用，其在生物网络中可能属于不同性质的群组，即更有可能成为“敌人”，基于局部二部结构在酵母 SL 网络中强烈富集[66]的认识，在两个基因间奇数长度的网络扩散路径数目的基础上定义了“奇扩散核”用于酵母 SL 对的分类任务，五折交叉验证下的 SVM 取得了受试者工作曲线(Receiver Operating Characteristic，ROC)下的面积(Area Under the Curve，*AUC*)为 0.897 的预测效果[67]. 而 Chipman 和 Singh[68]则借助网络扩散的思想将遗传相互作用与功能网络拓扑信息相结合，完成了不同生物数据之间的相互整合[68]. 具体来说，分别处于两个功能网络下的基因 i 和基因 j 是一对遗传相互作用，基因 i 与基因 j 的邻居基因 q 之间的拓扑相关性定义为基因 j 和基因 q 之间的拓扑相关性. 这种整合的逻辑类似于 Wong 等人[69]研究中提到的 2hop 特征(见“混合特征”部分)，Chipman 等人运用以上方法整合了酵母 PPI、GO 等网络，在决策树的五折交叉验证下取得了 0.969 的 *AUC* 值. Qi 等人与 Chipman 等人的两项工作表明网络上的动态特征在 SL 预测任务上的可用性，是网络特征分析和应用上的一个重要尝试.

3.3.2　纯生物功能特征

纯生物特征的提取从 SL 自身属性出发，更加关注 SL 作为一种特殊生物学现象的表达模式和表型差异，选择区别于非 SL 对的特殊生物功能、生物信号或其他潜在生物关联为特征参与监督学习过程，完成 SL 预测任务.

基因通过转录、翻译、修饰等一系列过程合成特定的蛋白质，蛋白质在特定细胞环境下折叠成特定的功能性三维结构，进而发挥特定的生物学功能和效应. 故一个基因的突变可能会导致其蛋白产品功能的丧失，进而产生不正常的代谢过程，最终影响生物表型. Li 和 Luo[31]认为这种蛋白质功能的丧失可能是由于蛋白产品中蛋白质结构域(对蛋白质的三级结构内独立折叠单元的描述)的丧失造成的，故假设蛋白质结构域与遗传相互作用之间存在很强的相关性. 并根据酵母的结构域信息对酵母基因互作编码：如果某对基因编码的两个蛋白质同时具有(单个具有/同时没有)特定的蛋白质结构域，则编码为

2(1/0).作者将蛋白质结构域特征作为 SVM 的输入用于酵母 SL 的预测任务上,具有较好的预测性能(AUC=0.925).

如"网络上的统计学方法"部分所说,SL 对被发现在人类细胞中会表现出不同于正常基因对的表达模式和突变情况,且 Jerby-Arnon 等人[26]和 Sinha 等人[44]人已经利用异常的表达和突变找到了一些可信的潜在 SL 对,这些研究表明癌症基因组信息在预测 SL 对上的作用.故 Lu 等人[28]在前人的研究上进一步分析人类癌症基因组中的 SCNA 和基因表达数据,发现无论是从基因表达水平还是 SCNA 水平上看,SL 对中的基因在癌症基因组中共同丢失的可能性比非 SL 要小得多,且 SL 对更有可能出现一个基因过表达而另一个低表达的表达模式,这种发现与 Jerby-Arnon 等人提出的算法依据一致.Lu 等人为了更加细致地刻画这种"共损失不足"的表达模式,定义了纯合子共缺失、杂合子共缺失、混合共缺失、共低表达、上下表达五种事件,并以它们的发生概率作为特征输入到由随机森林(Random Forest,RF)、SVM 等分类器构成的、基于均值集成的模型中,十折交叉验证的结果表明,预测人类 SL 的 AUC 可以达到 0.75.几乎是相似的尝试来自于 Das 等人[45],他们通过分析癌症患者的基因突变、基因表达、SCNA 和通路信息,量化了基因对之间的差异表达、表达相关性、互斥性和共享的路径,开发出了基于 RF 的 DiscoverSL 的 R 包.与前两项研究不同,Wu 等人[37]不再单独地观察某对基因的特殊生物信号,而是更加关注已知 SL 对与未知 SL 对之间的关系,认为未知对与已知 SL 对越相似,则越有可能具有相同的 SL 特征.Wu 等人集成了癌症基因组信息和蛋白质信息的多个分子网络,定义了基因对在基因表达谱、蛋白序列、PPI、通路关系、GO-BP、GO-CC 和 GO-MF 上的 7 个相似度,在 SNF 算法[70]的帮助下对以上相似度进行融合用于计算未知基因对与已知基因对间的总相似度,最终借助 k 近邻方法让与未知基因对最相似的 k 个已知 SL 对投票完成未知基因对的分类任务.

3.3.3 混合特征

在生物网络的拓扑特征提取时,会考虑分析网络节点及边的局部性质或网络全局结构,从这些网络的定量刻画中找出与 SL 对的关联关系.具有生物学意义的生物功能特征作为特征学习的重要补充可以提供宝贵的先验知识,在实际的 SL 预测过程中,两类特征混合使用可以取得更好的效果.

在机器学习刚刚被应用于 SL 基因对的预测任务时,人们对于 SL 关系的了解还不清晰,因此在进行特征提取任务时没有很强的目的性,通常会整合关于基因、染色体、mRNA、蛋白质、基因功能等各方面信息.Wong 等人[69]提出的 MNDT 方法从这些信息中提取了包括生物特征(如两基因的染色体距离、共同上游调控因子等)和网络特征(如聚类系数、共同邻居等)在内的共 123 个特征用于概率决策树上的酵母 SL 预测,模型的预测结果显示 SL 被正确预测的概率为 80%.值得注意的是,MNDT 方法中首次提出了 2hop 特征,用来描述两个基因 i、j 和第三个基因 q 之间的关系,如基因 i 与基因 q 间存在物理相互作用(P),基因 j 与基因 q 存在 SL 相互作用(S),则基因对(i,j)具有 2hop P-S.实验证明 2hop 特征在 SL 预测方面有着比一般生物特征更好的效果.MNDT 方法是当时整合多数据源预测遗传相互作用的最有效的方法,它的意义在于其首次严格证明了 SL 是可预测的,为后续的研究提供了广泛的参考.

MNMC 是基于多网络的多分类器方法[38],同 MNDT 方法一样,该方法整合了包括 PPI 网络、转录因子结合信息、基因功能注释信息、突变表型信息等来自基因组和蛋白组的多个网络,提取了 62 个单数据集特征(如基因的共表达、DNA 序列相似度等)和 90 个双数据集特征(通过两个网络的叠加形成)用于酵母 SL 对的预测,未知基因对是潜在 SL 对的得分来自各个分类器的属于 SL 类和非 SL 类的概率的乘积之间的差值,十折交叉验证的 AUC 结果约为 0.65.该方法的意义在于其拓展了 2hop 的概念,在 MNDT 中,只考虑两个基因之间是否存在通过第三个基因的 2hop 路径,即认为所有可能存在的路径拥有相同的权重,而在 MNMC 方法中并不只是尝试寻找任何这样的 2hop 路径,而是考虑最大的路径连接,选择构成叠加的两个分数的乘积的最大值赋为一对基因的新的特征值,这样的改进有利于提供更多的量化信息.然而,MNMC 结合多个分类器时没有考虑它们在预测性能上的差异,任何一个分类器性能的波动都会影响 MNMC 的整体预测结果.Wu 等人[71]在提出 MetaSL 方法时认识到这个问题,其设计的集成模型会根据不同模型在训练过程中的表现分配相应的权重,最终 SL 预测结果来自各分类器投票的加权共识.

最新的基于混合特征的方法是由 Benstead-Hume 等人[30]提出的 Slant 方法,这是一个基于 RF

的监督学习方法，它首次关注了 SL 在物种内和物种间的分类. 其物种内分类通过分别训练包括人类、果蝇等 5 个物种的 PPI 网络中提取到的三类特征（如度、特征中心性等节点特征；如共同邻居、最短路径等节点对特征；如共享的生物过程的注释数量等生物特征）来完成. Slant 使用从非人类 PPI 网络中提取的特征预测人类 SL 的 *AUC* 值普遍高于 0.65，当使用从人类 PPI 网络中提取的特征预测人类 SL 时，*AUC* 值更是达到了 0.965. 这说明 SL 关系在物种之间存在一定的保守性，即少数 SL 对在生物进化过程中可能不易发生变化. 此外，在这项研究中，Benstead-Hume 等人发现相比于非 SL 对，SL 对在 PPI 网络中共享更多的邻居、拥有更短的最短路径长度、让两个基因处于独立子图的代价更高，表明 SL 对在 PPI 网络上显著聚集.

基于网络特征学习预测人类 SL 关系的发展可以分为两个阶段. 研究初期受技术条件的限制，人们对于人类全基因组认识还不完善，只能借助低等真核生物（如酵母、线虫）的 SL 对来认识人类细胞中的潜在 SL 对. 随着高通量测序技术的发展，越来越多的基因组测序和临床数据的价值被发现，更多的分类任务直接作用在人类 SL 对上. 不可否认的是，基于低等真核生物的研究在预测 SL 任务上探索出了一条可靠的范式，这为基于人类细胞的研究提供了可靠的参考. 总的来说，这些基于机器学习建立的二元分类器，提取的大多数特征描述的是两基因间的结构相似性或功能相似性. 相比于功能相似性而言，PPI 网络上的结构相似性预测性能更好[38,68-69,71]，一种可能的解释是 PPI 网络本身就是一个功能网络，功能网络的结构相似性可能是结构和功能的双重结合[72]. 就分类器的选择来说，任何可用于分类任务的传统机器学习算法都可以在 SL 预测上被使用，但就研究结果来看 SVM 是使用最频繁的单分类器，也是效果最好的单分类器. 另外，集成分类器比单分类器有着更好的预测效果，RF 是预测任务中表现最好的传统机器学习算法. 对于论文中所描述的基于网络特征学习的方法的总结见表 5.

表 5　基于网络特征学习的方法

类别	方法	研究对象	算法描述	使用数据
纯网络拓扑特征	Paladugu[52]	酿酒酵母 SL	从酵母的 PPI 网络中提取一系列网络拓扑特征，用于 SVM 上预测泛癌背景下酿酒酵母的 SL. 首次证明 PPI 网络在 SL 预测任务上的强大能力，明确了 PPI 网络上含有丰富的基因信息.	PPI、SL
	Qi[67]	酿酒酵母 SL	根据两基因在基因互作网络上的奇数路径数定义“奇扩散核”，并用于 SVM 上完成酿酒酵母 SL 的预测任务.	SL、PPI
	Chipman[68]	酿酒酵母 SL	利用带重启的随机游走在各个功能性生物网络上计算基因对之间的拓扑相似度，并用于决策树分类器中预测酿酒酵母 SL. 方法证明了带重启的随机游走在生物网络上的使用价值比 2hop 特征高.	SL、GO、PPI
纯生物功能特征	Li[31]	酿酒酵母 SL	对蛋白质结构域信息编码并用于 SVM 中学习酿酒酵母 SL. 进一步证明蛋白质与基因相互作用之间的关联关系.	蛋白质结构域、SL
	Lu[28]	人类 SL（泛癌）	从癌症样本的 SCNA 和基因表达数据中发现 SL 对于非 SL 对之间的表达模式上的差异，并定义指标衡量以上的差异用于集成的多分类器之上预测泛癌背景下的人类 SL.	SL、SCNA、基因表达
	Das[45]	人类 SL（泛癌）	使用四个多组学特征（差异表达、表达相关性、互斥性和共享路径）的 RF 分类器来预测人类泛癌背景下的 SL. 进一步证明组学数据对于 SL 预测上的价值.	SL、基因突变、基因表达、SCNA 和基因通路
	Wu[37]	人类 SL（泛癌）	综合多种类型的基因对间相似性测度，采用 k-NN 算法实现基因对之间基于相似性的分类任务. 首次考虑未知 SL 对与潜在 SL 对之间的关系，突破了以往单独考虑 SL 对的桎梏.	基因表达、蛋白序列、PPI、基因通路和 GO
混合特征	MNDT[69]	酿酒酵母 SL	整合多个生物网络数据并在上整理特征用于概率决策树上学习并发现新的酿酒酵母 SL. 首次提出了生物网络上的 2hop 特征用于整合不同生物网络之间的关系，研究结果确定了哪些数据源对于 SL 预测来说是更加有效的，为后人的研究打下基础.	基因表达、PPI、基因序列、蛋白质序列、蛋白质复合物等
	MNMC[38]	酿酒酵母 SL	在多个生物网络上提取多个关于结构和功能的基因对特征用于集成分类器上酿酒酵母的 SL 的学习. 主要工作是对 Wong 等人[69]研究中没有考虑集成分类器的改进.	PPI、转录因子结合、基因功能注释等
	MetaSL[71]	酿酒酵母 SL	在多个生物网络上提取多个关于结构和功能的基因对特征用于集成分类器上酿酒酵母的 SL 的学习. 主要工作是对 Pandey 等人[38]研究中没有考虑集成分类器的各自性能的改进.	PPI、GO、基因表达、基因序列等
	Slant[30]	物种间 SL、物种内 SL	从 5 个物种的 PPI 网络中提取了各种基于图形的特征和其他生物特征，用于 RF 分类器，以预测物种内和物种间的 SL 对.	PPI、基因通路、SL

相比于网络上的统计学方法，基于网络特征学习的方法操作门槛较低，可整合的生物信息更加丰富；该类方法的研究历史最长，研究范式和研究技术都较为成熟，适合需要快速做出决策的场景. 但是，该类方法依赖于费时耗力且繁杂的特征工程，前期投入成本较大；另外，特征矩阵来源于多个特征的简单拼接，信息利用效率较低；最后，该类方法的可解释性也是一个值得深入研究的问题.

3.4 基于图表示学习的方法

基于网络特征学习的 SL 预测中，提取优质的网络特征往往可以得到较优的预测效果，特征的质量决定了预测任务的效率. 然而，对于某对基因从多个网络中分别提取特征后再拼接的做法受数据中噪声的影响较大，更重要的是，这样的做法不能捕获基因对在不同网络间的关联关系. 在网络研究领域，一些能够自动学习节点潜在特征的图表示学习方法[73-75]引起人们的关注，因其高效的数据集成能力被越来越多的用于生物信息学和生物医学的研究上[76]. 对于生物网络数据，采用图表示学习方法可以有效整合多个异构交互网络(如基因表达关系、基因交互关系等)，获得的节点向量化表示可以作为下游预测模型的输入完成二元分类(如图 2(b4)所示). 根据图表示方法的不同，我们将其分成以下四种：基于随机游走、矩阵分解、图神经网络和知识图谱的方法.

3.4.1 随机游走

在相互作用网络中具有相似拓扑角色或结构的蛋白质更有可能是功能相关的[7,10,18]，这种判断使我们能够通过相似的基因或蛋白质来推断未知基因或蛋白质的特性. 基于这种认识，随机游走常常用于生物网络上寻找相似节点.

带重启的随机游走(Random Walk with Restart, RWR)是被使用最多的游走方式，它强调节点在每次迭代时以概率 p_r 跳回种子节点. 其定义如下：

$$\boldsymbol{s}_i^{t+1}=(1-p_r)\boldsymbol{B}\boldsymbol{s}_i^t+p_r\boldsymbol{e}_i \tag{5}$$

其中 $\boldsymbol{B}$ 为转移概率矩阵(初始时为图的邻接矩阵的归一化形式)，p_r 表示 RWR 的重启概率；$\boldsymbol{e}_i$ 表示由节点 i 开始的随机游走的起点向量；$\boldsymbol{s}_i^t$ 表示由节点 i 开始的随机游走在 t 时刻的终点向量，其元素表示 t 步后每个节点的到达概率.

Cho 等人[32]提出的 Mashup 方法可以从多个异构数据类型的交互网络中学习基因的拓扑结构. 节点 i 的扩散状态 $\boldsymbol{s}_i$ 定义为由该节点开始的式(5)迭代后的稳态，则通过 RWR 可以得到特定网络 G 的扩散状态 $\boldsymbol{S}^{(G)}$，表示为网络中所有节点的拓扑向量；再对网络的向量表示降维，用低维向量 $\hat{\boldsymbol{S}}^{(G)}$ 来表示节点的拓扑环境：

$$\hat{S}_{ij}^{(G)}=\frac{\exp\{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{W}_j\}}{\sum_{j'}^{n}\exp\{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{W}_{j'}\}} \tag{6}$$

其中 $\boldsymbol{X}_i$ 和 $\boldsymbol{W}_i$ 是节点 i 的两个潜在向量，$\boldsymbol{X}_i$ 表示节点特征，$\boldsymbol{W}_i$ 表示上下文特征. 使用相对熵最小化 $\boldsymbol{S}^{(G)}$ 和 $\hat{\boldsymbol{S}}^{(G)}$ 之间的差距，则可以在特定网络上学习到所有节点特征 $\boldsymbol{X}$ 和每个网络的上下文特征 $\boldsymbol{W}$.

优化后的节点特征 $\boldsymbol{X}$ 可作为机器学习模型的输入，用于下游的推理或预测任务，文中将得到的节点特征用于 SVM 分类器上预测 SL 基因组合. Mashup 的意义在于提供了一个整合多个交互网络拓扑结构的工具，在基因相互作用预测、药物疗效预测等多个生物相关的预测任务上表现出适用性. 但其用于 SL 关系预测的结果表明，单纯的拓扑信息对于 SL 的认识是不够充分的，暗示了生物功能信息的重要性.

3.4.2 矩阵分解

矩阵分解技术已广泛应用于生物信息学相关的链路预测任务中，如 PPI 预测[77]、药物靶点预测[78-79]等. 矩阵分解通过矩阵补全学习基因之间的潜在关联，并在目标函数中设置约束来优化因子分解过程.

SL^{2}MF 是 Liu 等人[40]提出的基于逻辑矩阵分解[80]的概率模型方法，该方法根据 GO 注释和 PPI 网络中的拓扑属性计算基因之间的相似性并作为正则化信息，通过补全数据库 SynLethDB[24]中 SL 数据来完成预测任务. 逻辑矩阵分解的底层逻辑是通过贝叶斯公式最大化后验概率作为矩阵分解的损失函数 L_o：

$$L_o=p(O|\boldsymbol{U})\,p(\boldsymbol{U}|\sigma^2\boldsymbol{I}) \tag{7}$$

其中，O 为所有基因对的集合，$\boldsymbol{U}$ 为逻辑矩阵分解得到的潜在矩阵，$\boldsymbol{I}$ 为单位矩阵，σ^2 为控制高斯分布方差的参数，则 $p(O|\boldsymbol{U})$ 为观察到的 SL 数据的似然估计，$p(\boldsymbol{U}|\sigma^2\boldsymbol{I})$ 来源于零均值球面高斯先验[80]. 使用逻辑函数定义一对基因(i,j)是 SL 的概率为

$$P_{ij}=\frac{1}{1+\exp(-\boldsymbol{U}_i\boldsymbol{U}_j^{\mathrm{T}})} \tag{8}$$

其中 $\boldsymbol{U}_i$ 和 $\boldsymbol{U}_j$ 分别表示基因 i 和 j 的逻辑矩阵分解下的潜在向量.

此外，Liu 等人还考虑了 GO 注释和 PPI 网络的影响，将基因之间的 GO 语义相似度 s_{ij}^G[42]和 PPI 拓扑相似度 s_{ij}^P[53]作为标准来最小化两基因潜在向量上的差距，分别为 L 的第一项和第二项：

$$L=\sum_i\Big(\sum_{j\in N_i^G}s_{ij}^G\|\boldsymbol{U}_i-\boldsymbol{U}_j\|_2^2+\sum_{j\in N_i^P}s_{ij}^P\|\boldsymbol{U}_i-\boldsymbol{U}_j\|_2^2\Big) \tag{9}$$

其中，N_i^G 和 N_i^P 分别表示在 GO 网络和 PPI 网络中节点 i 的邻居节点集，$\|\boldsymbol{U}_i-\boldsymbol{U}_j\|_2^2$ 表示两个向量之间距离平方.

与 SL^{2}MF 类似，GRSMF[81] 同样选用 GO 相似度[42]作为反映基因相似性的先验信息，参与自表示矩阵分解(Self-representative Matrix Factorization，SMF)过程完成 SL 对的预测任务. GRSMF 的损失函数为

$$L=\min_{\boldsymbol{U}}\|\boldsymbol{X}-\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{U}\|_F^2+\beta\|\boldsymbol{U}\|_F^2+\beta R \quad (10)$$

其中第一项为 SMF 的目标任务，第二项为 L2 正则化项，第三项为 GO 相似度产生的正则化项.

与 SL^{2}MF 不同的是，GRSMF 依据观察到的 SL 相互作用，根据式(10)的表示规则学习基因之间的内部相似性. 而 SL^{2}MF 将基因投射到一个相空间，并根据其内积预测两个基因之间存在 SL 关系的概率，故 GRSMF 方法具有数据自适应性，避免了相空间维数等敏感参数的确定.

然而，SL^{2}MF 与 GRSMF 方法对于潜在矩阵的学习只能通过 GO 信息进行修正，对于信息更加丰富的基因突变和表达数据的处理却没有很好的融合方式. Liany 等人[82]为了在 SL 预测时可以整合更多的异构数据，进一步提高预测效率，提出三种方式拓宽了协同矩阵分解(Collaboration Matrix Factorization，CMF)[83]原本的使用范围. 前两种方式选择在矩阵分解之前对存在相同行列实体类型的矩阵进行转换(主成分分析和图特征)，第三种方式通过合并矩阵特定的权重来处理具有相同行和列实体类型的矩阵. 改进后的 CMF 模型可以在多个输入矩阵中包含相同实体类型时，学习到每个实体的唯一表示，这样的结果使得 CMF 用于 SL 预测任务时可以同时对多个描述基因关系的矩阵进行分解(例如描述患者和临床特征之间联系的临床数据，描述患者和基因之间联系的基因表达数据，描述物种和基因之间联系的系统发育数据等). CMF 的方法在多个不同分子层面的生物数据(如 SL、PPI、SCNA、mRNA 表达等)上被应用，3 个不同数据集上的 SL 预测结果证明，改进后的 CMF 相比于机器学习方法[28,32,38,71]、统计学方法[26-27]都显示出明显的优势.

3.4.3　图神经网络

图神经网络(Graph Neural Networks，GNN)通过深度神经网络模型学习输入网络中节点的低维向量表示，深度学习可以更好地捕捉输入和输出之间的非线性关系，使其能够识别数据背后的复杂模式. 因此，作为图数据上的全自动学习方式，GNN 已应用于各种生物信息学任务上.

EXP2SL 是 SL 预测任务上第一个使用神经网络结构的方法[34]，它首次考虑了 SL 在不同细胞系下的特异性表达，尝试开展了细胞系特异性遗传信息背景下的 SL 对的预测. 每个基因的细胞特异性表达信息来自于 L1000[47]的 978 维向量，经过编码层学习到基因的潜在表示，对于一对特定的基因，SL 的预测得分来自于两个基因潜在向量的线性表示. 为了提高模型性能，Wan 等人基于未标记样本和标记样本间的关联关系设计了贝叶斯个性化排名损失. 在 A549、A375 和 HT29 三个细胞系上的测试显示，半监督的神经网络结构为 EXP2SL 带来了高效和稳定的性能.

作为图神经网络用于 SL 预测任务上的初次尝试，EXP2SL 的成功表明了图神经网络的广阔发展前景. 但是，EXP2SL 简单地使用全连接网络结构作为编码层，其感受范围涉及输入的所有实体，故对于单个实体的潜在表示学习比较模糊，不能准确描述实际情况. 为了解决这样的问题，各种网络结构加入图神经网络的大家族中完成预测 SL 任务，在这里，我们从图卷积神经网络(Graph Convolution Networks，GCN)、图自编码器(Graph Auto-Encoders，GAE)和图注意力网络(Graph Attention Networks，GAT)三类上进行分别阐述.

(1) 图卷积神经网络

GCN 是图神经网络中的一种特殊网络结构，它强调通过聚合实体自身及其邻域的表示来产生实体的最终潜在表示. 依据邻域之间的依赖关系，可以有意识地控制卷积核的大小，故学习潜在表示的视野范围变小，更加精准的学习方式让潜在表示对特定实体更有区分度[73].

Cai 等人[84]将图卷积神经网络用于 SL 预测上，提出带有双 dropout 的图卷积神经网络方法 DDGCN. 借助 GCN，特定基因的描述信息可以同时来自该基因本身的表示和其邻居基因的表示. 为了解决普通 GCN 中标准 dropout 在稀疏图上的过拟合问题，Cai 等人提出粗粒度节点 dropout 和细粒度边缘 dropout，分别等价于同时从同一原始 SL 图中删除节点和边. 对于每一对基因，经过两个通道分别得到两个置信度得分，经过交叉熵优化的两个得分的几何平均数即为最终 SL 的预测得分，得分高的基因对越有可能被判定为 SL 对.

虽然 GCN 的运用为基因的潜在表示提供了更加准确的信息，但 Cai 等人的研究中只把 SL 数据作

为输入，缺乏其他生物网络的支持，单一数据源学得的模型可能存在偏倚.另外，使用 GCN 聚合邻居信息时会平等的考虑实体的每一个邻居，但在实际情况中，如同“近朱者赤，近墨者黑”描述的那样，不同邻居对于实体的影响效应或程度可能是完全不同的.为了解决这两个问题，研究人员又引入了 GAE 和 GAT.

(2)图自编码器

Cai 等人的研究中的“有偏”问题可以通过引入多视图(即建模时考虑多个数据源)来解决，但这样的设置可能会带来“噪声”，GAE 可以在一定程度上解决这个问题.GAE 分为编码和解码两个过程，通常情况下，GAE 使用 GCN 作为编码器，在编码器中输入图的拓扑信息和节点属性之后，以内积为解码器完成输入图重构[74]，我们期望目标输出和输入数据近似，故优化两者之间的误差可以完成模型的训练.

Lai 等人[54]提出的 MGE 方法集成了包括 PPI、GO、KEGG 在内的六个网络，通过 GCN 获得基因在不同生物网络背景下的嵌入表示，基因的最终表示根据不同生物网络下的潜在表示构造而成，取两个特定基因潜在表示的乘积应用 Sigmoid 得到预测该基因对间存在 SL 的概率，模型的训练通过 SL 图的重构完成.MGE 方法的有效性表明 GAE 在 SL 预测任务上由理论走向实践.

DDGCN 和 MGE 方法借助 GCN 或 GAE 都出色地完成了 SL 预测任务，但是仍有一定的问题存在.MGE 方法在整合多个数据源时没有考虑不同数据对于特定实体的影响作用，也没有考虑不同数据源中不同类型实体之间的相互关联；另外，MGE 方法仍旧没有解决 DDGCN 方法在聚合邻居信息时平等地看待每一个邻居的问题.

(3)图注意力网络

GAT 是 GCN 的进一步扩展，它强调在信息聚合时放大最重要数据的影响力[75].也就是说，GAT 在聚合某个特定实体的邻居信息时，会为该实体的每个邻居分配不同的权重；在聚合不同数据源的信息时，也可以针对不同数据类型学习不同的权重，注意力机制的添加弥补了 GCN 的不足.

Hao 等人[33]首次将 GCN、GAE 及 GAT 相结合，提出 SLMGAE 方法.同 MGE 方法一样，SLMGAE 方法的设计依照 GAE 的编码和解码原则，使用 GCN 对 PPI 拓扑相似度网络[53]、GO 语义相似度网络[42]、SL 三个网络分别编码，基因的潜在表示通过加权内积的解码器来完成.不同的是，SLMGAE 方法认为基因在 PPI 和 GO 两个支持视图下有着不一样的重要性，故重构支持视图时设计了注意力机制，即为每张图的重构损失分配一项归一化的权重，加权后的补充视图重构损失和 SL 图的重构损失共同构成模型训练的目标函数.

SLMGAE 方法尝试解决了不同数据类型聚合时的不足，但是仍没有解决不同邻居聚合时的问题.Long 等人[41]提出的 GCATSL 方法因为双注意力机制的存在，同时解决了以上两个问题.GCATSL 的双注意力机制包括节点级注意和特征级注意，节点级注意即在特定图 l 上聚合节点 i(其属性特征用 $\boldsymbol{h}_i$ 表示)的所有邻居 N_i^l 时，为每个邻居 j(其属性特征用 $\boldsymbol{h}_j$ 表示)分配特定的归一化权重 α_{ij}^l，其中 $\boldsymbol{W}$ 为模型的学习参数，$f(\cdot)$ 表示前馈神经网络：

$$e_{ij}^l = f(\boldsymbol{W}\boldsymbol{h}_i, \boldsymbol{W}\boldsymbol{h}_j) \tag{11}$$

$$\alpha_{ij}^l = \frac{\exp(e_{ij}^l)}{\sum_{t \in N_i^l} \exp(e_{it}^l)} \tag{12}$$

特征级注意即在对特定节点 i 的不同特征 F(来自于不同的数据源，其属性信息可以用 $\boldsymbol{z}_i^F$ 表述)聚合时，为每个特征分配特定的归一化权重 β_i^F，其中 $\boldsymbol{q}^{\mathrm{T}}$、$\boldsymbol{W}'$ 和 $\boldsymbol{b}$ 都为模型的学习参数：

$$w_i^F = \boldsymbol{q}^{\mathrm{T}} \cdot \tanh(\boldsymbol{W}'\boldsymbol{z}_i^F + \boldsymbol{b}) \tag{13}$$

$$\beta_i^F = \frac{\exp(w_i^F)}{\sum_{F=1}^{|F|} \exp(w_i^F)} \tag{14}$$

GCATSL 方法在聚合邻居信息时考虑了直接邻居和全局邻居(由 RWR 稳定后的得分矩阵确定的前 k 个节点)，聚合不同数据源时考虑了 PPI 拓扑相似度[53]网络和 GO 语义相似度[42]网络，在 3 个数据集上的大量实验结果表明，GCATSL 模型在预测 SL 相互作用方面优于传统机器学习方法、矩阵分解方法和部分神经网络方法.

3.4.4 知识图谱

知识图谱(Knowledge Graph，KG)是一种多关系图，可以用 $G(V,E)$ 表示，其中 V 表示知识图中的实体，E 表示知识图中的关系，对于一个头实体 b 和尾实体 t 之间的关系 τ 可以表示为三元组 $e=(b,\tau,t)$.KG 的优点在于可以对不同类型实体之间的多类关系建模[85].对于 SL 预测任务来说，可以将常用到的 PPI、GO、KEGG、SL 等网络整合为一张知识图，把基因之间的潜在关联直接引入图中，结合图神经网络的知识，可以获得更高质量的基因潜在

表示，将会为 SL 预测任务带来新的突破.

关于 KG 在 SL 预测任务上的最新尝试来自于 Wang 等人[25]提出的 KG4SL，该方法在 11 种实体的 24 种关系上建立了一个包含 54 012 个节点和 2 231 921 条边的知识图，该知识图中注入了 SL 预测中常用的数据源，如与基因相关的通路、GO-BP、GO-CC、疾病、药物等信息. KG4SL 方法融合了 GAE 和 GAT 的思想，首先会提取与输入基因对相关的实体和关系，建立特异性 KG 子图；考虑到每个实体与基因对的关系强度不同，KG 子图中的每条边都会分配不同的权重. 接下来，KG4SL 将 GAE 的编码和解码作为后续步骤，故预测得分来源于编码后基因潜在表示的内积. KG4SL 不仅在预测性能上带来了新的高度，更重要的是对癌细胞的复杂生物环境有了更加细致的建模，为 SL 预测研究提供了更全面的视角.

总的来说，图表示学习方法因为其高效的信息整合能力和较好的算法预测性能，成为目前 SL 预测上最有前景的方法类别. 随机游走倾向于全局查看图的拓扑结构来描绘实体所处的环境；矩阵分解认为实体的刻画可以通过间接地引入一个潜在因素来完成；图神经网络因为深度学习对于非结构化数据的强大表示能力而充满活力，知识图谱则以一种更加细粒度的方式来尽可能还原实体所处环境的真实情况. 因为知识图谱的图结构仍可以与图神经网络的方法相结合，故“知识图谱＋”将是 SL 预测任务上的一个重要发展趋势. 最后，不得不提的是，该类方法的可解释性仍旧是一个亟待解决的难题. 关于文中提到的所有基于图表示学习的方法的总结见表 6.

表 6　基于图表示学习的方法

类别	方法	研究任务	算法描述	使用数据
随机游走	Mashup[32]	人类 SL(泛癌)	通过学习基因的紧凑拓扑特征表示，集成了多个异构交互网络，并用 SVM 对预测人类 SL.	—
矩阵分解	SL^2MF[40]	人类 SL(泛癌)	采用逻辑矩阵分解学到的潜在矩阵对一对基因可能发生 SL 互作的概率建模.	PPI、GO、SL
	GRSMF[81]	人类 SL(泛癌)	提出了一种新的图正则化自表示矩阵分解模型用于 SL 交互预测.	GO、SL
	CMF[82]	人类 SL(泛癌)	提出三种方式用来对存在相同实体的生物关系建模，并用于人类 SL 预测任务.	PPI、基因共表达、SL、SCNA 等
图神经网络	EXP2SL[34]	人类 SL(特异性)	考虑细胞系特定的遗传信息，利用全连接网络从中学习 SL 间的潜在作用，并用于特定细胞系下 SL 预测.	细胞系特异性的基因表达、SL
	DDGCN[84]	人类 SL(泛癌)	利用图卷积神经网络对 SL 之间的关系建模，同时提出双 dropout 来解决数据稀疏的问题.	SL
	MGE[54]	人类 SL(泛癌)	在图自编码器的基础上提出一种多图集成的网络结构，在多个生物网络数据下预测基因对间的 SL 关系.	SL、PPI、GO、KEGG
	SLMGAE[33]	人类 SL(泛癌)	将图卷积神经网络、图自编码器、图注意力机制运用到模型设计中，在对多张网络数据的重构损失集成时，考虑了各自数据的贡献，进一步优化了 SL 预测.	SL、PPI、GO
	GCATSL[41]	人类 SL(泛癌)	提出图上下文注意网络，通过引入双注意力机制同时考虑了不同数据对于预测结果的贡献和不同邻居在信息聚合时的贡献，再一次优化了模型设计.	SL、PPI、GO
知识图谱	KG4SL[25]	人类 SL(泛癌)	在包含 11 个实体之间的 24 种关系的 SL 知识图上，融入图注意力网络和图自编码器的思想学习基因的潜在表示用于 SL 预测任务.	SynLethDB-SynLethKG

4　评估和验证

SL 预测研究的目的是发现可以用于癌症治疗的潜在药物靶点，模型预测结果的评估和验证是后续应用研究的基础. 在这里，我们认为 SL 预测结果的评估和验证可以从算法角度和应用角度两个方面来考虑.

4.1　算法角度

从算法角度上来说，研究者更加关心预测模型的算法性能，一种合理的推测是预测性能较好的模型理论上发现新药物靶标的能力也会更强.

为了评估模型的预测能力，k 折交叉验证经常被使用，具体来说，需要将数据平均分为 k 份，训练和测试过程重复 k 次，每次取 $k-1$ 份作为训练集，剩下的 1 份用于测试. 常用来衡量模型性能的指标有三组，分别为 ROC 曲线及 AUC、精确率-召回率曲线(Precision-Recall Characteristic，PR)及其曲线下面积($AUPR$)和敏感性($Sensitivity$)及特异性($Specificity$)[86]，通常来说，特定模型的预测性能会取 k 次测试结果的平均值.

SL 对的预测属于二分类任务，借助表 7 的混淆

矩阵可以方便且准确的刻画上述指标. 其中，TP 代表预测为 SL 对的 SL 对样本，FP 代表预测为 SL 对的非 SL 对样本，FN 代表预测为非 SL 对的 SL 对样本，TN 代表预测为非 SL 对的非 SL 对样本.

表 7 二分类混淆矩阵

混淆矩阵		真实值	
		正	负
预测值	正	TP	FP
	负	FN	TN

(1) ROC 曲线与 AUC 值

ROC 曲线可以用来展示一个分类模型在所有分类阈值下的表现，其横坐标是假阳性率(False Positive Rate，FPR)，即所有负样本中预测结果为正样本的比例；ROC 曲线的纵坐标是真阳性率(True Positive Rate，TPR)，即所有正样本中预测结果为正样本的比例. ROC 曲线下的面积定义为 AUC 值，AUC 值越高，表示模型在预测任务上的表现越好. FPR、TPR 和 AUC 的定义如下：

$$FPR=\frac{FP}{FP+TN} \tag{15}$$

$$TPR=\frac{TP}{TP+FN} \tag{16}$$

$$AUC=\frac{\sum_{i\in PositiveClass} rank_i-\frac{(TP+FN)\times(TP+FN+1)}{2}}{(TP+FN)\times(FP+TN)} \tag{17}$$

其中，$rank_i$ 表示第 i 个样本的编号，样本编号是由获得的预测概率得分升序排列得到的.

(2) PR 曲线与 $AUPR$ 值

PR 曲线对于不平衡类别的预测结果有更为妥当的衡量，其横坐标是召回率($Recall$)，指在所有正样本中被正确预测出来的比例；其纵坐标是精确率($Precision$)，指在所有样本中预测结果与真实结果一致的比例. PR 曲线下的面积定义为 $AUPR$ 值，$AUPR$ 值越高，表示模型在预测任务上的表现越好. $Recall$ 和 $Precision$ 的定义如下：

$$Recall=\frac{TP}{TP+FN} \tag{18}$$

$$Precision=\frac{TP}{TP+FP} \tag{19}$$

(3) 敏感性与特异性

敏感性指的是所有正样本中预测结果为正样本的比例；特异性指的是所有负样本中预测为负样本的比例. 敏感性高，则越容易鉴定出目标，即漏诊率低；特异性高，则越不容易误报，即误诊率低. $Sensitivity$ 和 $Specificity$ 的定义如下：

$$Sensitivity=\frac{TP}{TP+FN} \tag{20}$$

$$Specificity=\frac{TN}{FP+TN} \tag{21}$$

本文中提到的所有 SL 预测方法性能的总结见表 8(表中不包括“网络上的统计学方法”和“基于网络结构变化的方法”，因为暂时不能定量的描述这些方法的性能).

表 8 SL 预测方法性能的总结

方法	验证策略	AUC	$AUPR$
Paladugu[52]	十折交叉验证		0.4400
Qi[67]	五折交叉验证	0.8970	
Chipman[68]	五折交叉验证	0.9690	
Li[31]	五折交叉验证	0.9250	
Lu[28]	十折交叉验证	0.7500	
MNMC[38]	十折交叉验证	0.6500	
MNDT[69]	四折交叉验证	0.8500	
Wu[37] ($k=19$)	十折交叉验证	0.8500	0.8600
MetaSL[71]	五折交叉验证	0.8710	
Slant[30]	五折交叉验证	0.9650	
Mashup[32]	十折交叉验证		0.5900
SL^{2}MF[40]	五折交叉验证	0.8480	0.2388
GRSMF[81]	五折交叉验证	0.9230	
CMF[82]	三折交叉验证	0.9550	0.9430
EXP2SL(HT29)[34]	五折交叉验证	0.9690	0.8800
DDGCN[84]	五折交叉验证	0.8782	0.3442
MGE[54]	五折交叉验证	0.9553	0.9555
SLMGAE[83]	五折交叉验证	0.9174	0.9428
GCATSL[41]	五折交叉验证	0.9375	0.9483
KG4SL[25]	五折交叉验证	0.9470	0.9564

4.2 应用角度

从应用角度上来说，研究者不仅关注 SL 预测模型的性能，还应关注预测模型的新预测结果在生物实践中的可用性，这里的生物实践包括生物信息学分析和新预测任务. 本文重点介绍常用于 SL 预测研究中的三种生物信息学分析，包括基因富集分析[87]、药物敏感性分析[88]和患者生存分析[27]. 另外，本文还会就新预测任务对现有研究进行简单回顾.

(1) 基因富集分析

基因富集分析来源于超几何分布，计算在 $L+U$ 个物品中抽出 l 个物品，成功抽出 u 个指定种类的物品的概率. 在生物信息学中，前文的概率称为富集，更具体的说是 u 个物体在 U 个物体上的富集.

在 SL 预测任务中，我们关心的问题是：在人类全基因组的尺度上，参与 SL 关系的基因是不是在癌症相关的生物学功能上显著富集. 因此，富集分析用来将预测得到的 SL 对的基因按照基因功能注释

信息分类，从而观察基因在某类特定功能上的富集是不是有统计学意义，基因富集分析解决了“存在 SL 关系的基因更倾向于参与哪些生物功能”的问题. 根据现有研究可知，SL 基因对之间存在功能冗余关系[7,19,89]，因此，期望预测的 SL 对之间共享更多的功能注释信息. 通过将 GO/KEGG pathway 的注释信息匹配到单个基因上，还可以获得关于基因对之间共享的基因功能的认识. 当预测的 SL 对比非 SL 对显著共享更多的注释信息时，则可以认为 SL 的预测结果存在一定的合理性.

(2) 药物敏感性分析

药物敏感性分析可以查看 SL 作用是否对常见抗癌药物治疗存在正向促进作用，该分析建立在两个基因同时突变对细胞产生不利影响而单独突变却不会对细胞生长造成影响的 SL 定义之上. 对于任意一对 SL 基因对(G_A,G_B)来说，若基因 G_A 可以作为某药物的靶标，那么，当分别向基因 G_B 存在突变和不存在突变的两类细胞系(分别记为 Mutant 和 Wild)中施加该药物，因为 SL 作用的存在，期望观察到 Mutant 细胞系中的一半细胞死亡所需的药物浓度(Half-maximal Inhibitory Concentration，IC50)比 Wild 细胞系更低. Mutant 细胞系对药物的耐受程度低暗示了基因 G_A 和基因 G_B 间 SL 作用激活后的抑制癌细胞生长的负作用的存在.

完成药物敏感性分析需要确认每一对 SL 基因组合中的靶标基因，并将已有细胞系按照靶标基因的 SL 伙伴是否发生突变分为 Mutant 组和 Wild 组，再针对两组中细胞系的 IC50 值做显著性检验，最后可以根据检验结果查看，在特定癌症药物下特定 SL 基因组合对于细胞生长是否存在消极影响. 最后，需要说明的是，药物敏感性分析中所需的数据可以在抗癌药物敏感性基因组学数据库(Genomics of Drug Sensitibity in Cancer，GDSC)[90]中获得.

(3) 患者生存分析

患者生存分析来源于时序检验(log-rank 检验)，可以在一定的时间范围内估计生存时间和变量之间的关系，生存分析的结果是生存曲线，其横坐标是随访时间，纵坐标是累计生存概率，曲线上的每一点表示在当前随访时间点下所有随访患者的生存率，因此生存曲线描述的是一段时间内一组患者的生存状况.

在 SL 预测研究中，生存分析中的变量为患者的基因相互作用网络中是否存在 SL 关系，当 SL 作用存在时，两个基因在细胞中同时发生突变，对癌细胞会有致命的作用. 此时，癌症患者可以借助 SL 关系抵抗癌症的进一步发展，延长生存时间，改善生存状态. 因此，我们期望得到的结果是存在 SL 关系的患者的生存时间比不存在 SL 关系的患者的生存时间显著更长. 值得一提的是，在大多数研究中，通常以两基因同时突变或共低表达来模拟 SL 关系的发生. 最后，生存分析需要的患者临床和基因突变数据都可以从 TCGA 数据库中获得.

(4) 新预测任务

验证预测结果合理性的另一种办法是将 SL 预测结果用于与癌症相关的新预测任务中，考虑到目前关于这部分的研究较少，暂时难成体系. 因此，本文主要以回顾相关研究为主.

Cheng 等人[91]借助 GTEx[92]中的正常组织的基因表达数据，量化了非癌人体组织中 SL 基因对的共失活水平，作为对癌症发展抵抗的衡量(称为 SL 负载，SL 负载越高，对于癌症发展的抵抗力越强)，在比较了正常组织和癌症组织中的 SL 负载后发现，SL 负载越高的人面临的癌症风险越低，癌症的发病年龄越大，即 SL 关系出现得越频繁越有可能延缓癌症的发生. Lee 等人[93]选择将预测得到的 SL 对预测不同癌症患者对于不同靶向药物的反应情况，发现 SL 在患者表达数据中所占的比例与患者的药物反应显著相关，这种相关性在更多的情况下表现为 SL 占比越大，患者对于癌症靶向药物的反应越好，暗示了 SL 治疗癌症的可能性.

5　结　语

“合成致死”概念的出现为癌症的精准靶向治疗注入了新的活力，为了找出更多潜在的癌症治疗靶点，人们经历了从实验鉴定到计算机筛选的过程. 实验鉴定 SL 对依赖于大规模高通量筛选技术的发展，然而实验的高成本和脱靶频发等问题使得体外 SL 对检测变得困难重重. 在实验积累的大量生物数据的支持下，基于计算机的硅筛选得到了迅速的发展，本文特别关注基于生物网络的 SL 预测方法，从四个角度分别综述了现有的主要方法，为后续的 SL 预测研究提供了重要参考.

目前，基于生物网络的 SL 预测研究正处于快速发展阶段，未来必定有着广阔的发展前景，我们认为未来研究面临的挑战和可能的潜在发展方向主要在以下几个方面：

(1) 数据的极度不均衡性. 受制于筛选技术和 SL

作用本身的特异性，一对没有被鉴定为 SL 的基因对并不能确定它们一定是非 SL 对，故 SL 的标签数据存在着少量的正样本和大量的未标记样本[7,94]. 本文中提到的大多数方法的负样本来源于未标记样本的随机采样，但是通常未标记样本中包含正样本，错误的标签指定会导致错误的样本分类. 因此，我们认为未来的预测方法需要考虑如何将正样本和无标签学习(PU 学习)，即“启发式的从未标记样本中找到可靠的负样本”[95]用到 SL 的预测上来，推动 SL 预测的进一步发展.

(2) 数据质量. SL 预测研究中使用的正样本只有很少一部分来自于低通量实验筛选的可靠数据，更多的数据来源是高通量实验筛选技术. 高通量实验筛选受样本质量、筛选文库、数据质量控制等相关因素的影响，其筛选结果的可靠性相比于低通量筛选大打折扣，最后用于 SL 预测研究上的数据质量不可避免的会存在假阳性问题. 另外，常用于 SL 预测研究中的 PPI 等其他生物数据是不完整且“有偏”的，已有研究证明使用“有偏”数据会产生“有偏”结果[48]. 故 SL 预测研究急需一个金标准的 SL 正负样本集合，且生物数据的发展也会推动 SL 预测研究的进步.

(3) SL 知识图谱研究. SL 的预测以不同分子层面的生物知识为起点展开，现存的 SL 网络作为主要信息，PPI、GO 等其他生物网络作为辅助信息，根据挖掘的各个网络之间的相互关联做出合理的生物预测. KG4SL[25]的成功已经明确了知识图谱在揭示实体相互关联上的重要性，未来基于知识图谱的研究应该持续向前发展，进一步丰富基于知识图谱的 SL 预测算法.

(4) 单细胞测序数据的深度挖掘. 经常作为补充信息参与预测过程的基因转录组数据来源于 RNA 测序技术，其得到的是一群细胞转录组的平均数据，这忽略了单个细胞的特异性信息，如单个基因的特异性表达. 而 SL 的组织特异性一直是一个难以处理的棘手问题，在不同器官或组织上，同一对基因往往会呈现出不同的 SL 反应结果. 如果能在分析 SL 的过程中结合单细胞数据，则可以获得细胞的异质性信息，也许会加深我们对于肿瘤微环境的认识，为我们提供一个思考问题的全新维度.

(5) SL 预测算法的可解释性. 本文提到的所有预测算法类型中，网络上的统计学方法和基于网络结构变化的方法是可解释性相对较强的两类. 但是，当机器学习和深度学习用于 SL 预测之后，模型的优化开始变得盲目，越来越多的学习模型致力于提高模型的预测精度，而忽略了模型因为“黑盒”性质带来的解释性差的问题. 这种做法是不理智的，我们认为未来的预测模型应该将“减法”融入模型开发中，模型的可解释性应该是模型设计的重点工作之一.

(6) SL 预测结果的评估和验证. 现有的 SL 预测研究的目标是基于 SL 关系发现新的抗癌药物靶点. 但是，大多数研究的终点是借助五折交叉验证来证明预测算法的性能，只有很少的研究会通过功能富集分析、药物敏感性分析等来间接验证其新预测结果的合理性，在研究最后与生物学家合作开展新药物靶标验证工作的研究更是屈指可数. 由此可见，对于 SL 预测结果的评估和验证是一个关键而复杂的任务[94]，目前仍然没有一个系统和权威的方法出现，希望随着这项工作的深入，可以尽快解决这一难题.

(7) SL 概念的延申. 在已知所有的研究中，只有 Apaolaza 等人[63]注意到在多个基因之间存在的高阶负遗传相互作用. 我们认为高阶 SL 作用应该引起重视，因为基因参与特定的生物功能或代谢过程时更有可能形成团簇结构[96]. 当把 SL 的概念延申至多个基因之间时，可能会发现新的药物靶标，并且这些靶标可能与联合用药有关. 另外，当处理高阶相互作用时，超图的引入可能会带来新的研究课题.

(8) 特异性 SL 的预测. 在本文提到的所有研究中，只有少数几个研究[34,39,59]的预测任务建立在特定的癌症背景下或特定的细胞系中，这是因为目前可用的特定遗传背景的数据资源或数据量有限. 如在 Wan 等人[34]的研究中，来自 A375 和 HT29 细胞系的正样本只有 18 对，较小的数据量上很难发现稳定的统计学规律，监督学习过程也很容易发生过拟合. 因此，受现有数据的影响，特异性 SL 的预测一直处于困境之中.

参 考 文 献

[1] Siegel R L, Miller K D, Fuchs H E, et al. Cancer statistics, 2022. CA: A Cancer Journal for Clinicians, 2022, 72(1): 7-33

[2] Ciardiello F, Tortora G. EGFR antagonists in cancer treatment. New England Journal of Medicine, 2008, 358(11): 1160-1174

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

[4] Liu C, Zhao J, Lu W, et al. Individualized genetic network analysis reveals new therapeutic vulnerabilities in 6700 cancer genomes. PLoS Computational Biology, 2020, 16(2): e1007701

[5] Bridges C B. The origin of variations in sexual and sex-limited characters. American Naturalist, 1922, 56(642): 51-63

[6] Dobzhansky T. Genetics of natural populations; Recombination and variability in populations of Drosophila pseudoobscura. Genetics, 1946, 31(3): 269-290

[7] O'Neil N J, Bailey M L, Hieter P. Synthetic lethality and cancer. Nature Reviews Genetics, 2017, 18(10): 613-623

[8] Qiao Hong-Xia, Huang Guo-Xiang, Huang Ying-Hui, et al. Application of synthetic lethality in the precise oncology. Chinese Science Bulletin, 2018, 63(12): 1123-1129(in Chinese)
(乔红霞，黄国翔，黄映辉等. SL 在精准肿瘤学中的应用. 科学通报，2018, 63(12): 1123-1129)

[9] Farmer H, McCabe N, Lord C J, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature, 2005, 434(7035): 917-921

[10] Hartwell L H, Szankasi P, Roberts C J, et al. Integrating genetic approaches into the discovery of anticancer drugs. Science, 1997, 278(5340): 1064-1068

[11] Tong A H, Evangelista M, Parsons A B, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science, 2001, 294(5550): 2364-2368

[12] Ooi S L, Pan X, Peyser B D, et al. Global synthetic-lethality analysis and yeast functional profiling. Trends in Genetics, 2006, 22(1): 56-63

[13] Lehner B, Crombie C, Tischler J, et al. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nature Genetics, 2006, 38(8): 896-903

[14] Collins S R, Roguev A, Krogan N J. Quantitative genetic interaction mapping using the E-MAP approach. Methods in Enzymology, 2010, 470: 205-231

[15] Hannon G J. RNA interference. Nature, 2002, 418(6894): 244-251

[16] Wang T, Wei J J, Sabatini D M, et al. Genetic screens in human cells using the CRISPR-Cas9 system. Science, 2014, 343(6166): 80-84

[17] Conde-Pueyo N, Munteanu A, Solé R V, et al. Human synthetic lethal inference as potential anti-cancer target gene detection. BMC Systems Biology, 2009, 3(1): 1-15

[18] Srivas R, Shen J P, Yang C C, et al. A network of conserved synthetic lethal interactions for exploration of precision cancer therapy. Molecular Cell, 2016, 63(3): 514-525

[19] Brough R, Frankum J R, Costa-Cabral S, et al. Searching for synthetic lethality in cancer. Current Opinion in Genetics & Development, 2011, 21(1): 34-41

[20] Zhang F, Wu M, Li X J, et al. Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. Journal of Bioinformatics and Computational Biology, 2015, 13(3): 1541002

[21] Lu X, Kensche P R, Huynen M A, et al. Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. Nature Communications, 2013, 4(1): 1-10

[22] Liu C, Ma Y F, Zhao J, et al. Computational network biology: Data, models, and applications. Physics Reports, 2020, 846: 1-66

[23] Li X, Mishra S K, Wu M, et al. Syn-Lethality: An integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. BioMed Research International, 2014, 2014: 196034

[24] Wang J, Wu M, Huang X, et al. SynLethDB 2.0: A web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. Database, 2022, 2022: baac030

[25] Wang S, Xu F, Li Y, et al. KG4SL: Knowledge graph neural network for synthetic lethality prediction in human cancers. Bioinformatics, 2021, 37(Supplement_1): i418-i425

[26] Jerby-Arnon L, Pfetzer N, Waldman Y Y, et al. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell, 2014, 158(5): 1199-1209

[27] Lee J S, Das A, Jerby-Arnon L, et al. Harnessing synthetic lethality to predict the response to cancer treatment. Nature Communications, 2018, 9(1): 2546

[28] Lu X, Megchelenbrink W, Notebaart R A, et al. Predicting human genetic interactions from cancer genome evolution. PLoS One, 2015, 10(5): e0125795

[29] Shen J P, Zhao D, Sasik R, et al. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nature Methods, 2017, 14(6): 573-576

[30] Benstead-Hume G, Chen X, Hopkins S R, et al. Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. PLoS Computational Biology, 2019, 15(4): e1006888

[31] Li B, Luo F. Predicting yeast synthetic lethal genetic interactions using protein domains//Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine. Washington, USA, 2009: 43-47

[32] Cho H, Berger B, Peng J. Compact integration of multi-network topology for functional analysis of genes. Cell Systems, 2016, 3(6): 540-548

[33] Hao Z, Wu D, Fang Y, et al. Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder. IEEE Journal of Biomedical and Health Informatics, 2021, 25(10): 4041-4051

[34] Wan F, Li S, Tian T, et al. EXP2SL: A machine learning framework for cell-line-specific synthetic lethality prediction. Frontiers in Pharmacology, 2020, 11: 112

[35] Zamanighomi M, Jain S S, Ito T, et al. GEMINI: A variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. Genome Biology, 2019, 20(1): 1-10

[36] Ye H, Zhang X, Chen Y, et al. Ranking novel cancer driving synthetic lethal gene pairs using TCGA data. Oncotarget, 2016, 7(34): 55352

[37] Wu L L, Wen Y Q, Yang X X, et al. Synthetic lethal interactions prediction based on multiple similarity measures fusion. Journal of Computer Science and Technology, 2020, 36(2): 261-275

[38] Pandey G, Zhang B, Chang A N, et al. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Computational Biology, 2010, 6(9): e1000928

[39] Yang C, Guo Y, Qian R, et al. Mapping the landscape of synthetic lethal interactions in liver cancer. Theranostics, 2021, 11(18): 9038-9053

[40] Liu Y, Wu M, Liu C, et al. SL2MF: Predicting synthetic lethality in human cancers via logistic matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 17(3): 748-757

[41] Long Y, Wu M, Liu Y, et al. Graph contextualized attention network for predicting synthetic lethality in human cancers. Bioinformatics, 2021, 37(16): 2432-2440

[42] Wang J Z, Du Z, Payattakool R, et al. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, 23(10): 1274-1281

[43] Costanzo M, VanderSluis B, Koch E N, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science, 2016, 353(6306): aaf1420

[44] Sinha S, Thomas D, Chan S, et al. Systematic discovery of mutation-specific synthetic lethals by mining pan-cancer human primary tumor data. Nature Communications, 2017, 8(1): 1-13

[45] Das S, Deng X, Camphausen K, et al. DiscoverSL: An R package for multi-omic data driven prediction of synthetic lethality in cancers. Bioinformatics, 2019, 35(4): 701-702

[46] Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas(TCGA): An immeasurable source of knowledge. Contemporary Oncology, 2015, 19(1A): A68-77

[47] Subramanian A, Narayan R, Corsello S M, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 2017, 171(6): 1437-1452

[48] De Kegel B, Quinn N, Thompson N A, et al. Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. Cell Systems, 2021, 12(12): 1144-1159

[49] Kranthi T, Rao S B, Manimaran P. Identification of synthetic lethal pairs in biological systems through network information centrality. Molecular Biosystems, 2013, 9(8): 2163-2167

[50] Prasad T K, Goel R, Kandasamy K, et al. Human protein reference database—2009 update. Nucleic Acids Research, 2009, 37(suppl_1): D767-D772

[51] Magen A, Sahu A D, Lee J S, et al. Beyond synthetic lethality: Charting the landscape of pairwise gene expression states associated with survival in cancer. Cell Reports, 2019, 28(4): 938-948

[52] Paladugu S R, Zhao S, Ray A, et al. Mining protein networks for synthetic genetic interactions. BMC Bioinformatics, 2008, 9: 426

[53] Chua H N, Sung W K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 2006, 22(13): 1623-1630

[54] Lai M, Chen G, Yang H, et al. Predicting synthetic lethality in human cancers via multi-graph ensemble neural network// Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Mexico, Guadalajara, 2021: 1731-1734

[55] Szklarczyk D, Gable A L, Nastou K C, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 2021, 49(D1): D605-D612

[56] Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science, 2021, 30(1): 187-200

[57] Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Research, 2019, 47(D1): D330-D338

[58] Bhinder B, Shum D, Djaballah H. Comparative analysis of RNAi screening technologies at genome-scale reveals an inherent processing inefficiency of the plasmid-based shRNA hairpin. Combinatorial Chemistry & High Throughput Screening, 2014, 17(2): 98-113

[59] Liany H, Jeyasekharan A, Rajan V. ASTER: A method to predict clinically actionable synthetic lethal genetic interactions. bioRxiv, 2021: 2020.10.27.356717

[60] Orth J D, Thiele I, Palsson B Ø. What is flux balance analysis. Nature Biotechnology, 2010, 28(3): 245-248

[61] Megchelenbrink W, Katzir R, Lu X, et al. Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(39): 12217-12222

[62] Tobalina L, Pey J, Planes F J. Direct calculation of minimal cut sets involving a specific reaction knock-out. Bioinformatics, 2016, 32(13): 2001-2007

[63] Apaolaza I, José-Eneriz S, Tobalina L, et al. An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. Nature Communications, 2017, 8(1): 1-9

[64] Aye Y, Li M, Long M J C, et al. Ribonucleotide reductase and cancer: Biological mechanisms and targeted therapies. Oncogene, 2015, 34(16): 2011-2021

[65] Apaolaza I, San José-Enériz E, Valcarcel L V, et al. A network-based approach to integrate nutrient microenvironment in the prediction of synthetic lethality in cancer metabolism. PLoS Computational Biology, 2022, 18(3): e1009395

[66] Ye P, Peyser B D, Spencer F A, et al. Commensurate distances and similar motifs in genetic congruence and protein

interaction networks in yeast. BMC Bioinformatics, 2005, 6(1): 1-13

[67] Qi Y, Suhail Y, Lin Y, et al. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. Genome Research, 2008, 18(12): 1991-2004

[68] Chipman K C, Singh A K. Predicting genetic interactions with random walks on biological networks. BMC Bioinformatics, 2009, 10(1): 1-11

[69] Wong S L, Zhang L V, Tong A H, et al. Combining biological networks to predict genetic interactions. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101: 15682-15687

[70] 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

[71] Wu M, Li X, Zhang F, et al. In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. Cancer Informatics, 2014, 13(Suppl 3): 71-80

[72] Li J R, Lu L, Zhang Y H, et al. Identification of synthetic lethality based on a functional network by using machine learning algorithms. Journal of Cellular Biochemistry, 2019, 120(1): 405-416

[73] Wang H, Wang J, Wang J, et al. GraphGAN: Graph representation learning with generative adversarial nets//Proceedings of the AAAI Conference on Artificial Intelligence. California, USA, 2018: 2508-2515

[74] Kipf T N, Welling M. Variational graph auto-encoders. arXiv: 1611.07308, 2016

[75] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv:1701.10903, 2018

[76] Liu C, Han Z, Zhang Z K, et al. A network-based deep learning methodology for stratification of tumor mutations. Bioinformatics, 2021, 37(1): 82-88

[77] Wang H, Huang H, Ding C, et al. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. Journal of Computational Biology, 2013, 20(4): 344-358

[78] Liu Y, Wu M, Miao C, et al. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Computational Biology, 2016, 12(2): e1004760

[79] Ezzat A, Zhao P, Wu M, et al. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016, 14(3): 646-656

[80] Johnson C C. Logistic matrix factorization for implicit feedback data. Advances in Neural Information Processing Systems, 2014, 27(78): 1-9

[81] Huang J, Wu M, Lu F, et al. Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization. BMC Bioinformatics, 2019, 20(Suppl 19): 657

[82] Liany H, Jeyasekharan A, Rajan V. Predicting synthetic lethal interactions using heterogeneous data sources. Bioinformatics, 2020, 36(7): 2209-2216

[83] Singh A P, Gordon G J. Relational learning via collective matrix factorization//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2008: 650-658

[84] Cai R, Chen X, Fang Y, et al. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. Bioinformatics, 2020, 36(16): 4458-4465

[85] Liu Qiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques. Journal of Computer Research and Development, 2016, 53(3): 582-600(in Chinese)
(刘峤，李杨，段宏等. 知识图谱构建技术综述. 计算机研究与发展，2016，53(3)：582-600)

[86] Lever J. Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. Nature Methods, 2016, 13(8): 603-605

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

[88] Zheng S, Aldahdooh J, Shadbahr T, et al. DrugComb update: A more comprehensive drug sensitivity data repository and analysis portal. Nucleic Acids Research, 2021, 49(W1): W174-W184

[89] Hartman IV J L, Garvik B, Hartwell L. Principles for the buffering of genetic variation. Science, 2001, 291(5506): 1001-1004

[90] Yang W, Soares J, Greninger P, et al. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research, 2012, 41(D1): D955-D961

[91] Cheng K, Nair N U, Lee J S, et al. Synthetic lethality across normal tissues is strongly associated with cancer risk, onset, and tumor suppressor specificity. Science Advances, 2021, 7(1): eabc2100

[92] Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. Nature Genetics, 2013, 45(6): 580-585

[93] Lee J S, Nair N U, Dinstag G, et al. Synthetic lethality-mediated precision oncology via the tumor transcriptome. Cell, 2021, 184(9): 2487-2502

[94] Madhukar N S, Elemento O, Pandey G. Prediction of genetic interactions using machine learning and network properties. Frontiers in Bioengineering and Biotechnology, 2015, 3: 172

[95] Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Machine Learning, 2020, 109(4): 719-760

[96] Guo Mao-Zu, Wu Xue-Jian, Zhao Ning, et al. A method for mining core modules of cancer based on multi-omics biological network. Scientia Sinica Informationis, 2017, 47(11): 1510-1522(in Chinese)
(郭茂祖，武雪剑，赵宁等. 一种基于多组学生物网络的癌症关键模块挖掘方法. 中国科学：信息科学，2017，47(11)：1510-1522)

LIU Chuang, Ph. D., professor. His research interests focus on complex networks and bioinformatics.

SHU Sheng-Li, M. S. candidate. Her research interests focus on complex networks and bioinformatics.

ZHAN Xiu-Xiu, Ph. D., associate professor. Her research interests focus on machine learning and complex networks.

ZHANG Zi-Ke, Ph. D., professor. His research interests focus on complex intelligent computing and computational propagation.

Background

Synthetic lethality (SL) describes a specific biological phenomenon between two genes on the basis of which successful precedents have been achieved in the development of cancer drugs, and recently, the cancer control situation has made the task of finding the existence of synthetic lethal gene combinations more and more urgent. Generally, the search for synthetic lethal pairs includes bio-experimental identification and bioinformatics-based computer prediction. Bioinformatics-based computer prediction, as an important complement to experimental identification, can make full use of biological information at different molecular levels to detect the gene combination, which can greatly improve the detection efficiency.

SL is of great significance for explaining complex biological processes and clinical diagnosis and treatment of cancer. Therefore, it is an important direction for computational biology research to use the massive high-throughput data to mine and predict SL pairs from a computational perspective by constructing data analysis models and computational methods. Biological networks are logically organized forms that connect individual molecular units and are important data samples in bioinformatics research. Most of the prediction work on synthetic lethal pairs is built on them, and this paper focuses on these methods and classifies them into four categories: statistical methods on networks, methods based on network structure variation, methods based on network feature learning, and methods based on graph representation learning. After reviewing the latest progress of related prediction models and studies, we compare the algorithmic ideas, application scenarios, and advantages and disadvantages of each type of methods in detail. In addition, several challenges faced in SL prediction research are further summarized, and the future development direction is targeted, hoping to provide some useful references and ideas for future related research.

This study is an exploratory study belonging to the analysis of cancer mechanism and antitumor drug target prediction research based on biological network modeling, which helps to deepen the understanding of synthetic lethality relationship. By sorting out the related prediction methods and current difficulties faced, we clarify a specific direction of drug target prediction based on synthetic lethality and confirm that the specific prediction of synthetic lethality is a focus of the research. In fact, we have made some preliminary attempts to improve the prediction of synthetic lethal pairs after considering multi-omics biological information.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61873080 and 92146001), the Major Project of the National Social Science Fund of China (Grant No. 19ZDA324) and the Fundamental Research Funds for the Central Universities.