湖南第一师范学院

Research and Application of Algorithms Based on Protein-protein Interaction Networks

基于蛋白质相互作用网络的算法研究及其应用

汤希玮◎著

科学出版社

基于蛋白质相互作用网络的算法研究及其应用

汤希玮 著

科学出版社

北京

内 容 简 介

为了揭示蛋白质网络的动态特征，酵母的时间序列基因表达谱被用于分离静态的蛋白质网络，进而构建随时间变化的动态蛋白质网络。为了理解蛋白质的关键性和聚集性，基因表达谱和亚细胞位置信息被引入蛋白质相互作用网络中，从而设计了一系列蛋白质复合物挖掘算法和关键蛋白质识别算法，并对新提出的算法进行了详细的、多角度的比较测试。考虑到蛋白质网络在疾病基因识别过程中所起的巨大作用，本书的后半部分，从蛋白质网络出发，重点研究了各种疾病基因与蛋白质网络的关系，提出了一系列与肿瘤等复杂疾病有关的基因识别算法。最后，集中探讨了蛋白质网络研究的新方向。

本书适合计算机科学、生物信息学、生物医学研究人员和教师阅读使用，也可作为相关专业本科生和研究生的学习参考材料。

图书在版编目(CIP)数据

基于蛋白质相互作用网络的算法研究及其应用/汤希玮著. —北京：科学出版社，2018. 8

ISBN 978-7-03-058571-4

Ⅰ. ①基… Ⅱ. ①汤… Ⅲ. ①蛋白质-基因组-研究 Ⅳ. ①Q51

中国版本图书馆 CIP 数据核字(2018)第 196893 号

责任编辑：胡庆家 / 责任校对：邹慧卿

责任印制：张 伟 / 封面设计：无极书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京建宏印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 8 月第 一 版 开本：720×1000 B5

2018 年 8 月第一次印刷 印张：13 1/2 插页：1

字数：180 000

定价：98.00 元

（如有印装质量问题，我社负责调换）

前　言

自2010年接触蛋白质-蛋白质相互作用网络以来，作者在该领域的研究已经持续8年了，研究的方向涵盖动态蛋白质网络、关键蛋白质、蛋白质复合物和疾病基因等，在生物信息学相关的主要国际期刊上发表了一系列文章，有了这些研究积累之后就准备出一部专著，一方面是梳理并总结自己这些年的研究成果，另一方面是给生物信息学研究人员提供积极的借鉴，希望能够提高其研究工作效率。

生物体是一个非常复杂的系统。为了理解生命活动的运行机制，有必要从系统级别进行研究。随着高通量技术的飞速发展，产生了海量的组学数据，这给网络科学在生物信息学中的应用打下了坚实的数据基础。研究人员因此构建了各种类型的生物学网络，如代谢网络、基因调控网络、转录网络和蛋白质相互作用网络等，尽管这些网络普遍存在“假阳性”和“假阴性”，但是在大数据时代，数据的完整性远比数据的准确性更重要，况且生物网络很好地满足了从整体上研究生命体的客观需求，因此基于生物网络的应用研究一直深受广大科研人员的青睐。

当前各种蛋白质组学数据库中搜集的蛋白质-蛋白质相互作用网络都是静止的，并没有从时间和空间上展开，但是生物体本身是活跃的，基因可能在不同的时间和空间（器官组织）上被表达，为了深刻理解细胞系统的动态性，有必要构建时间序列的蛋白质网络以及不同组织器官的蛋白质网络（空间网络）。作者利用时间序列的基因表达谱将酿酒酵母的静态蛋白质相互作用网络分离为36个不同时刻的动态蛋白质网络，该研究有一定的开创性，但也只是起了抛砖引玉的作用，作者认为利用时间或空间特征明显的生物学数据将静态的人类蛋白质相互作用网络转化为动态网络，应该是未来该领域最有价值的研究方向。

在蛋白质-蛋白质相互作用网络中，存在一个核心的蛋白质集合，这些蛋白质数量不多，但是在生物体的生命活动中起决定性作用，它们构

成了生命体的基座，如果敲除这样的蛋白质会造成生命体不可逆转的损毁，这些蛋白质就是所谓的关键蛋白质。早期的关键蛋白质研究专注于网络的拓扑特征，后来为了克服组学数据中存在的假阳性，人们开始整合不同来源、不同类型的生物学数据，构建加权蛋白质网络并设计算法侦测关键蛋白质。作者将基因表达数据和蛋白质亚细胞位置数据引入蛋白质网络，先后提出了三种关键蛋白质侦测算法，这些算法有助于加深生物医学科学家对关键蛋白质的认识。作者认为由于关键蛋白质和疾病基因之间具有密切的关系，因此，关键蛋白质信息对识别疾病基因具有重要的意义，在设计计算机方法识别疾病基因的过程中，如何利用关键蛋白质数据是未来需要解决的重要科学问题。

聚集性是蛋白质-蛋白质相互作用网络最明显的特征，在大规模蛋白质相互作用网络环境中，少部分重要的、两两相互作用并紧密结合在一起的蛋白质构成了一个个的蛋白质复合物。蛋白质复合物产生的各种分子机制能执行大量生物功能，是生命活动中许多生物过程得以实现的基础。后基因组时代最大的挑战之一就是从蛋白质网络中识别蛋白质复合物。作者整合不同来源的生物学数据，设计了两种算法从加权蛋白质-蛋白质相互作用网络中识别蛋白质复合物。该研究从图论的角度出发，将蛋白质网络映射成图，图中的结点代表蛋白质，图中的边表示蛋白质之间的相互作用关系，密度子图对应网络中的复合物，挖掘蛋白质复合物的问题转化为识别图中的密度子图。考虑到生物网络的动态性，仅仅从静态网络中挖掘蛋白质复合物，还不足以反映生命活动的本质特征，作者认为未来的研究应该从挖掘静态的蛋白质复合物转向识别反映生物体动态特性的功能模块。

由于“关联推定”(guilt-by-associate)原则的存在，使得基于蛋白质-蛋白质相互作用网络识别疾病基因成为可能。一般的算法是以已知的疾病基因为种子结点，根据关联推定原则，在网络中寻找与种子结点密切相关的结点，这些结点最有可能是新的疾病基因。作者基于网络局部特征和全局特征提出了两种疾病基因识别算法。

蛋白质网络的动态性、关键性和聚集性等特征，为识别疾病基因开辟了新的研究道路，本书将这四个方面的研究结合为一个有机整体，力

图展现蛋白质网络研究的核心内容，但由于个人水平有限，不足之处还请读者批评指正。

特别感谢中国国家自然科学基金委员会资助了本书的出版（项目编号：61472133 和 61772089）。

汤希玮

2018 年 5 月

目　录

第1章 绪 论

1.1 蛋白质网络的计算分析

1.1.1 蛋白质网络及其研究所面临的挑战

因为人类基因组测序已经实现[1,2]，所以遗传学领域现在站到了重要的理论和实践进步的门槛上。这使得全面理解某生物体编码的蛋白质的表达、功能和调控变得至关重要，从而也诞生了蛋白质组学。蛋白质组是指在某一时刻被基因组、细胞、器官或生物体表达的蛋白质的完整集合，也是指在确定的条件下，在某一给定的时刻和给定类型的细胞或生物体中被表达的蛋白质的集合。蛋白质组学系统研究蛋白质的各种属性，目的是详细地描述在健康和疾病状态下，生物系统的结构、功能和控制方式。过去十年，在生物信息学领域中，关于蛋白质组学的研究呈爆发式增长。

在蛋白质组学中，一个特别受关注的领域是蛋白质相互作用的本质和它在生命活动中所扮演的角色。当两个或更多的蛋白质绑定在一起时，它们之间就以相互作用的方式实现其生物功能。蛋白质-蛋白质相互作用调控着大量的生物过程，包括转录的激活与抑制，免疫的、内分泌的和药理学的信号，细胞与细胞之间的相互作用，以及代谢和发育控制等[5-8]。蛋白质-蛋白质相互作用在生物中起着各种作用，并因相关的组成、亲缘关系和生存时间的不同而不同。侧链残基之间的共价键关联性是蛋白质折叠、装配和相互作用的基础[9]。这些关联性使蛋白质之间与之内的各种相互作用及关联变得更容易。根据不同的结构和功能特征，蛋白质-蛋白质相互作用可以按照几种不同的方式分类[10]。根据它们的相互作用的外观，可分为相似低聚物或相异低聚物的相互

作用;根据它们的稳定性,可分为必须的或非必须的相互作用;根据它们的持续性,可分为瞬时的或永久的相互作用。一个给定的相互作用可能属于这三种类别中任意一种,也可能需要在某一条件下再次重新分类。例如,某相互作用可能在活的有机体内是瞬时的,但是在某一细胞条件下又变成永久的了。

研究者通过分析被注释的蛋白质,发现涉及同一细胞过程的蛋白质往往彼此之间发生相互作用[11]。根据未知蛋白质与功能已知的目标蛋白质之间的相互作用,研究者能假定未知蛋白质的功能。绘制蛋白质相互作用地图不仅有利于深刻理解蛋白质的功能,而且使为了解释细胞过程的分子机制而进行的功能路径的建模变得容易。蛋白质-蛋白质相互作用的研究对理解细胞内的蛋白质怎样活动很重要。在一个给定的细胞蛋白质组中,特征化蛋白质的相互作用将是沿着理解细胞的生物化学过程的道路上的里程碑。

两个或更多的蛋白质和某一特定的功能目标相互作用的结果能被几种不同的方式验证。研究者概括了蛋白质-蛋白质相互作用的几种可衡量的影响[12]。

· 蛋白质-蛋白质相互作用能改变酶的动力学属性,这可能是变构效应(allosteric effect)级别或底物结合(substrate binding)级别的微妙改变的结果。

· 蛋白质-蛋白质相互作用充当了允许底物通道(substrate channeling)的共性机理。

· 蛋白质-蛋白质相互作用创造了新的结合部位,尤其是对小的效应分子而言。

· 蛋白质-蛋白质相互作用使蛋白质失去活性或毁灭。

· 通过和不同的绑定伙伴相互作用改变蛋白质的基底(substrate)的特异性。

为了适当地理解蛋白质-蛋白质相互作用在细胞中的重要性,人们需要识别不同的相互作用,理解它们在细胞中是怎样产生的,确定相互作用的效果。

近年来,高通量实验技术(如双杂交系统[13,14]、质谱分析方法[15,16]

和蛋白质芯片技术[17-19])使得蛋白质相互作用数据日益丰富。研究者已经从这些异构的数据源构建了综合的蛋白质-蛋白质相互作用网络。然而,当前可用的大量蛋白质-蛋白质相互作用数据对实验研究提出了挑战。为了理解没有被特征化的蛋白质,蛋白质-蛋白质相互作用网络的计算分析变成了必要的补充工具。

蛋白质-蛋白质相互作用网络能被描述为相互作用的蛋白质构成的复杂系统。蛋白质-蛋白质相互作用网络的计算分析从蛋白质-蛋白质相互作用网络结构的表示开始。最简单的表示是包含结点和边的数学图形[20]。蛋白质代表图形中的结点,物理上相互作用的两个蛋白质表示为由一条边连接的邻居结点。基于这种图形表示,研究者能设计各种计算机方法(如数据挖掘、机器学习和统计方法)以揭示蛋白质-蛋白质相互作用网络在不同级别上的组织性。通过对网络的图形形式的检查,人们能获得各种深刻的见解。例如,图形中相邻的蛋白质通常被认为有共同的功能(即"guilt-by-association"特性)。因此,蛋白质的功能可能通过着眼于与它相互作用的蛋白质和它所属的蛋白质复合物而被预测。另外,网络中紧密相连的子图可能形成在某一生物过程中作为一个单元起作用的蛋白质复合物。网络拓扑特征(例如,它是否是无标度的小网络或者受幂次定律支配)的研究也能提高作者对生物系统的理解[21]。

一般而言,蛋白质-蛋白质相互作用网络的计算分析是具有挑战性的。研究者面临的常见的主要困难如下:

·蛋白质-蛋白质相互作用是不可靠的。大规模的实验产生了大量的假阳性。例如,据文献[22]报道,高通量的酵母双杂交(Y2H)实验大约50%是可靠的。在当前研究的其他蛋白质-蛋白质相互作用网络中可能存在许多假阳性。

·一个蛋白质可能有几种不同的功能。一个蛋白质可能被包含在一个或多个功能组中。因而,重叠的聚类应该从蛋白质-蛋白质相互作用网络中识别出来。因为传统的聚类方法通常产生没有交集的聚类,它们不能有效地运用于蛋白质-蛋白质相互作用网络。

·具有不同功能的两个蛋白质频繁地相互作用。在不同的功能组

中，蛋白质之间的随机连接扩大了蛋白质-蛋白质相互作用网络的拓扑复杂性，给侦测明确的分区部分（聚类或复合物）带来了困难。

最近，复杂系统[21,23]的研究试图从拓扑的角度，理解和特征化那些系统的结构性能。研究者已经在复杂系统中观察到小世界效应[24]、无标度分布[25,26]和层次模块性[27]。因而拓扑方法能被用于在一定程度上解决以上提到的挑战，并使蛋白质-蛋白质相互作用网络的有效的和精确的分析变得容易。

1.1.2 蛋白质网络研究的具体内容

（1）拓扑特征分析

为了量化网络中一个被选结点有 k 个链接的概率，研究者[26]提出了度分布（表示为 $P(k)$）的概念。不同类型的网络以度分布为特征。例如，随机网络服从泊松分布（Poisson distribution）。相反，无标度网络服从幂律分布，即 $P(k)\sim k^{-\gamma}$。幂律分布表明少数 Hub 结点与大部分结点相连。当 $2\leqslant\gamma\leqslant3$ 时，Hub 结点在网络中起了重要作用[26]。最近的文献[28-31]指出，蛋白质-蛋白质相互作用网络具有无标度网络的特征，因此这种网络的度分布近似于幂律分布。在无标度网络中，大部分蛋白质仅参与了很少的相互作用，然而，少部分 Hub 结点参与了大量相互作用。

蛋白质-蛋白质相互作用网络也具有所谓的“小世界效应”，即任意两个结点可能通过一条仅含有少数链接短路径连接在一起。小世界效应最初是在社会学研究中作为一个概念提出来的[32]，它是一系列网络（包括互联网[21]、科研合作网[33]、英语词典[34]、代谢网络[35]和蛋白质-蛋白质相互作用网络[36,31]）的特征。尽管小世界效应是随机网络的属性，但是无标度网络中的路径长度比小世界效应预测到路径长度短得多[37,38]。因而，无标度网络是超小型的。短路径长度表明在代谢物中的局部扰动可能很快扩散到整个网络。在蛋白质-蛋白质相互作用网络中被高度连接的结点（即 Hub 结点）之间很少直接彼此相连[39]。这不同于社会网络的配对的性质，在社会网络中有大量人际关系的个体之间往往直接有联系。相反，生物网络具有异配的属性，网络中被高度连

接的结点之间很少有关联。

许多研究者最近提出了结点中心性、网页排名、聚类系数、中介中心性和桥接中心性等中心性指标作为衡量网络中部件的重要性手段[40-45]。例如，中介中心性[46]用于侦测划分一个网络时的最优位置[47,48]。研究者建议将改进后的介数割(betweenness cut)方法用于整合基因表达信息后的加权蛋白质-蛋白质相互作用网络[49]。也有人建议将结点度作为识别关键的网络部件的重要基础[29]。在这种模型中，幂律网络对随机攻击而言很强健，但是对有目标的攻击却显得非常脆弱[50]。有研究团队识别了三个真核蛋白质-蛋白质相互作用网络中的关键和非关键基因之间的(酵母、蠕虫和果蝇)度、介数、密切度中心性之间的区别[51]。新的子图中心性指标也被提出，它能特征化每个结点在一个网络的所有子图中的参与情况[42,52]。研究者通过弧形删除识别了致命结点，从而使网络子部件的隔离更容易[53]。文献[54]设计了聚类方法识别代谢路径中的功能模块，并且按照与被侦测功能模块相关的每个部件的拓扑位置对路径中的每个部件的角色进行了分类。

(2) 模块性分析

文献[6]引入了功能模块的理念，并提供了系统地分析生物网络的大部分概念性工具。蛋白质-蛋白质相互作用网络中的功能模块代表了功能相关的蛋白质的最大集合。换句话说，它是由互相关联于某一给定生物过程或功能的蛋白质组成。大量聚类方法已经被用于蛋白质-蛋白质相互作用网络中的识别功能模块。然而，这些方法在精度方面受到限制，因为存在不可靠的相互作用而且网络本身是复杂的[22]。尤其是模块的重叠模式以及模块之间的交叠所导致的蛋白质-蛋白质相互作用网络的拓扑复杂性给功能模块的识别带来了挑战。因为蛋白质通常在不同的环境中执行不同的生物过程或功能，所以真实的功能模块是有重叠的。另外，不同功能模块之间的频繁的动态的交叉连接是有生物意义的，必须被考虑[55]。

为了分解复杂性，生物网络中模块的分层结构被提了出来[27]。这种模型的架构是基于具有嵌入式模块性的无标度拓扑结构。在这种模

型中，少数 Hub 结点重要性被强调，这些结点被视作在网络被干扰期间物种生存的决定性因素以及分层结构的重要骨架。这种分层网络模型用于蛋白质-蛋白质相互作用网络可能是合理的，因为细胞的功能本质上是分层的，而且蛋白质-蛋白质相互作用网络包括少数具有生物致命性的 Hub 结点。

识别蛋白质-蛋白质相互作用网络中的功能模块或者模块性分析可能通过聚类分析成功地实现。在解释网络部件之间相互关系以及网络的拓扑结构方面，聚类分析是很有价值的。典型地说，聚类方法致力于侦测蛋白质-蛋白质相互作用网络的图形表示中紧密相连的子图。例如，极大团算法[56]用于侦测充分相连的完整的子图。为了弥补这种算法导致的高密度临界点的不足，研究者通过使用密度阈值或者最优化目标密度函数，能够识别相关的密度子图而不是完整子图[56,57]。许多使用可选密度函数的基于密度的聚类算法已经被提出[58-60]。

正如前面所指出的那样，分层聚类方法能用于生物网络，因为功能模块具有分层的本质[27,61]。这类方法迭代地合并结点或者递归地将图划分为两个或更多子图。为了迭代地合并结点，两个结点或者两个结点集合之间的相似性或距离需要被测量[62,63]。图的迭代划分涉及将被分割的结点或边的选择。基于划分的方法也被用于生物网络。例如，受限的邻居搜索聚类算法就是这样一个基于划分的方法[64]，它使用成本函数确定最佳划分(聚类)[65]。还有研究团队基于共享更多共同邻居数的蛋白质对具有更高的相似性的这一规律，利用统计方法聚类蛋白质复合物[66]。

拓扑指标能结合进蛋白质-蛋白质相互作用网络的模块性分析中。研究人员发现蛋白质-蛋白质相互作用网络中识别的桥接结点能充当蛋白质模块之间的连接结点，因而移除桥接结点能保留网络的结构完整性。这些研究在蛋白质-蛋白质相互作用网络的模块性分析中起了非常重要的作用。移除桥接结点能产生一组来自于网络的不相连的部件。因此，在评估蛋白质-蛋白质相互作用网络中的模块的位置和数目时，使用桥接中心性移除桥接结点可能是一个出色的预处理步骤。这一研究[67,68]的结果显示，与其他方法相比，桥接结点移除方法能产生更大的

功能模块，预测功能模块的精度也越高。

(3) 蛋白质功能预测

因为蛋白质功能预测本身是蛋白质-蛋白质相互作用网络分析的最终目标，所以蛋白质功能预测一直是这一领域的研究热点。

尽管大量关于酵母的研究已经进行，但是，在酵母的数据库中仍然存在大量在功能上没有被特征化的蛋白质。人类蛋白质的功能注释信息为全面理解细胞机制提供了坚实的基础，并且对药物发现和开发很有价值。蛋白质-蛋白质相互作用网络的可用性以及人们对它的研究兴趣促进了预测蛋白质功能的计算机方法的发展。

模块化的算法能预测蛋白质的功能。如果一个未知的蛋白质包含在一个功能模块中，那么它可能对模块所代表的功能有贡献。产生的功能模块可能因此提供一个预测未知蛋白质功能的框架。每个产生的模块可能包含少量未被特征化的蛋白质和大量被特征化的蛋白质。可以假定未知的蛋白质在模块的功能实现方面起到积极的作用。然而，因为模块化过程的精度很低，所以通过模块化手段预测蛋白质功能的精度也不高。为了获得更高的可靠性，应该直接从蛋白质-蛋白质相互作用网络的连接性或拓扑特征入手预测蛋白质的功能。

一些基于拓扑结构预测蛋白质功能的方法已经出现。最简单的是邻居计数方法，它通过直接邻居的蛋白质的已知功能的频率预测未知蛋白质的功能[55]。直接邻居的大部分功能也能在统计上进行评估[69]。如果考虑一个蛋白质的直接邻居的功能，那么其功能就可以假定为独立于所有其他蛋白质。这种假定导致了马尔科夫随机场模型的出现[70,71]。最近，已知蛋白质和未知蛋白质的共同邻居的数量已经被作为功能预测的基础[72]。

机器学习也广泛用于分析蛋白质-蛋白质相互作用网络，尤其是预测蛋白质功能。研究者开发了各种基于不同信息源的方法预测蛋白质功能。这些方法使用的输入信息包括蛋白质结构和序列、蛋白质结构域、蛋白质-蛋白质相互作用、遗传相互作用和基因表达分析。预测的精度因使用了多数据源信息而得到提高。基因本体(gene ontology，GO)数据库[73]就是那样一个语义集成的例子。

(4) 动态蛋白质网络

细胞在时间和空间上的变化活动对其生存和繁殖起着重要的作用。细胞的动态属性隐含在构成细胞生理基础的蛋白质网络的拓扑结构中。例如,裂殖酵母中的细胞分裂就受到它的蛋白质网络的控制。生物学家研究生物网络的动态性已经很多年了,一般而言,他们都关注受限环境中的单个的基因或蛋白质以及特定的相互作用。在更大一点的规模上,生物学家已经绘制了跨物种的代谢网络,它内在地包含了时间信息并依赖特定代谢物[74]。最近几年,研究者遇到了从全局角度研究动态蛋白质网络的前所未有的机会,因为各种各样的高通量实验数据使得他们能从基因组规模上理解分子相互作用。从而,许多研究者提出了大量定量模拟和仿真动态蛋白质网络系统的计算机方法[75-85]。还有许多研究团队关注利用基因组规模的实验数据集分析网络的动态性。

(5) 数据融合

近年来,整合不同来源的各种生物学数据来提高计算机方法精度的研究越来越受研究人员的青睐,并成为了蛋白质网络研究的热点。

正如前面提到的那样,网络的高度复杂性及其所包含的虚假的连接可能影响了计算机方法的精度。传统的计算机方法仅仅能预测两个蛋白质是否共享某一特定的功能,但是不能预测它们共享的全部功能。算法的有效性因为没有考虑蛋白质功能的全部可用信息而大打折扣。通过整合其他生物数据源,研究者能改善这些方法的可靠性。

· 融合 GO 信息

在生物信息学社区中,GO 是当前最广泛和提取的最好的本体数据库。它致力于解决基因及其产物的一致性描述。GO 数据库包括 GO 术语及其相互关系。前者是良好定义的生物术语,这些术语被组织成三个通用的概念分类:生物过程(biological process,BP)、分子功能(molecular functions,MF)和细胞组分(cellular component,CC)。GO 数据库也给每一个 GO 术语提供了注释,每个基因可以被注释为一个或更多的 GO 术语。因而 GO 数据库及其注释是发现功能信息的重要资源。最初,GO 信息被用于使基于表达数据的分析变得容易[86-88],后来,它被