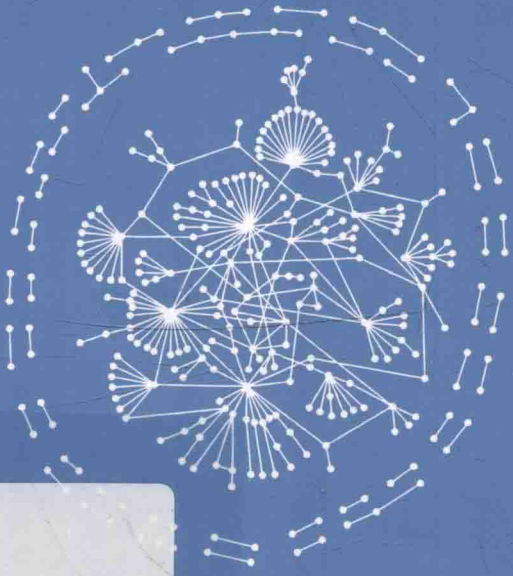
基于随机游走模型的蛋白质网络研究

JIYU SUIJI YOUZOU MOXING DE DANBAIZHI WANGLUO YANJIU

彭玮 著



云南大学出版社
YUNNAN UNIVERSITY PRESS

基于随机游走模型的蛋白质网络研究

JIYU SUIJI YOUZOU MOXING DE DANBAIZHI WANGLUO YANJIU

彭玮 著

云南大学出版社
YUNNAN UNIVERSITY PRESS

图书在版编目（CIP）数据

基于随机游走模型的蛋白质网络研究 / 彭玮著. --
昆明 : 云南大学出版社, 2017
ISBN 978-7-5482-3005-2

Ⅰ. ①基… Ⅱ. ①彭… Ⅲ. ①蛋白－基因组－研究
Ⅳ. ①Q51

中国版本图书馆CIP数据核字(2017)第116591号

策划编辑：赵红梅
责任编辑：周　飞　石　可
封面设计：王嫣一

基于随机游走模型的蛋白质网络研究

JIYU SUIJI YOUZOU MOXING DE
DANBAIZHI WANGLUO YANJIU

彭玮　著

出版发行：云南大学出版社
印　　装：云南南方印业有限责任公司
开　　本：787mm×1092mm　1/16
印　　张：8.25
字　　数：168千
版　　次：2017年8月第1版
印　　次：2017年8月第1次印刷
书　　号：ISBN 978-7-5482-3005-2
定　　价：28.00元

社　　址：昆明市一二一大街182号（云南大学东陆校区英华园内）
邮　　编：650091
电　　话：（0871）65033244　65031071
E－mail：market@ynup.com

前　言

生物信息学是生命科学和计算机科学相结合形成的一门新学科。大多数生命活动都是由多个蛋白质相互作用共同完成的。随着高通量实验技术的发展产生了大量的蛋白质相互作用数据。通过构建蛋白质相互作用网络使得我们能从系统水平来理解分子生物学系统，进而研究分子功能及其相互作用，同时也为生物进化研究提供了新的视角。随机游走模型在图上已有深入的研究。通过在图中的随机游走可以很方便地获得图中节点间的关联关系，也可以用来研究整个图的结构。

本书介绍了应用随机游走模型，结合蛋白质相互作用网络数据和其他生物数据，来解决关键蛋白质识别、蛋白质复合物挖掘、蛋白质功能预测以及蛋白质保守复合物挖掘几个问题的方法。

本书不仅对关键蛋白质识别、蛋白质复合物挖掘、蛋白质功能预测以及蛋白质保守复合物挖掘这几个生物信息学问题进行了深入研究，而且也介绍了随机游走模型在生物网络中的成功应用。本书可以为本领域研究者提供翔实的、专业的和独创的知识，具有很高的参考价值。

本书由彭玮（昆明理工大学）执笔，受到国家自然科学基金（31560317，61502214）、云南省自然科学基金（2016FB107）的资助。

由于时间仓促，书中欠妥和纰漏之处在所难免，恳请读者和同行不吝指正。

目　录

第一章 绪 论

1.1 蛋白质相互作用网络及随机游走模型

1.1.1 蛋白质间的相互作用

细胞是生物体组成最基本的单位。所有生物无论是细菌还是人类，虽然它们的细胞结构和组织复杂程度不相同，但是都有一个共同的细胞机制。那就是它们都经历一个从 DNA 转录到 RNA 再合成蛋白质的过程。DNA 是遗传信息的携带者。编码在 DNA 序列中的遗传信息通过 DNA 的复制传递给后代，同时通过转录成信使 RNA（mRNA）分子，然后通过信使 RNA 翻译生成体内的各种蛋白质。蛋白质是细胞生命的重要组成部分。它们参与细胞生命的各个过程。蛋白质是用于构建细胞结构的主要成分，也可以作为酶来催化体内的各种新陈代谢，还在信号传导、免疫反应、细胞粘附以及细胞周期方面起到了重要的作用。因此研究蛋白质的结构和功能对于揭示生命的奥秘意义重大。

然而，研究表明细胞的生命活动并不是通过单个蛋白质完成的，而是由一个或多个蛋白质聚集在一起来共同完成一个指定的功能。很多在细胞生命活动中重要的分子过程，比如 DNA 的复制，都是通过大量的蛋白质相互作用来共同完成的。蛋白质间的相互作用在生命活动的各个方面都会存在。比如，一个细胞外部的信号是通过跟信号传导有关的蛋白质之间的相互作用传导到细胞内部的，这个过程叫做信号传导，在很多生物过程和疾病（如癌症）中起着重要的作用。蛋白质间的相互作用也可能长期存在，从而形成蛋白质复合物。也可能是一个蛋白质运载另外一个蛋白质从而发生相互作用。比如通过 importin α/β 蛋白质运转系统，经过核孔将物质从细胞质传输到细胞核或者从细胞核传输到细胞质。或者一个蛋白质仅仅是因为要改变另外一个蛋白质而与它发生短暂的相互作用。比如蛋白质催化酶磷酸化修饰目标蛋白质。这种蛋白质的修饰本身也会改变蛋白质间的相互作用。例如，一些具有 SH2 结构域的蛋白质只有当它们的氨基酸、酪氨酸被

磷酸化时，才与其他蛋白质结合。总而言之，蛋白质间的相互作用几乎遍及细胞活动的每一个过程。这些相互作用的信息，有助于我们理解生命活动的本质，提高我们对疾病的认识，并为提供新的治疗方案打下良好的基础。

目前识别蛋白质相互作用的方法主要分为三大类。第一类是利用计算的方法系统地识别蛋白质间的相互作用，其中一个例子是 I. Rodriguez-Llorente 等[1]构建的中华根瘤菌（Sinorhizobium meliloti）共生关系网络。这项研究识别了 263 个新的与共生关系有关的蛋白质相互作用，并且利用该网络的拓扑特性作为指导，通过实验技术来验证新的涉及共生不同阶段的蛋白质。第二类方法就是利用亲和标记物（affinity tag）或者取下大块的蛋白质复合物，然后利用质谱技术来识别其中的蛋白质[2]。利用这类方法可以识别一些相互作用子网。比如说识别大肠杆菌（E. coli）中细胞被膜相关的蛋白质及它们的相互作用[3]。第三类方法是通过改变某些基因，然后分析表型来观察基因改变后的影响，从而预测蛋白质的相互作用。这类方法中的酵母双杂交系统（Y2H）[4,5]是最常用的识别蛋白质相互作用的方法。近年来，基于以上实验和计算方法产生了大量的不同物种的蛋白质相互作用数据，其中涉及物种达 1 000 多种，而且这些数据都存储在公开的数据库中，研究者可以很方便地下载。

1.1.2 蛋白质相互作用网络的研究内容

目前基于蛋白质相互作用数据的研究主要是构建蛋白质相互作用网络，网络中的节点表示的是蛋白质，边表示的是蛋白质之间的相互作用。然后采用图表的方法来对蛋白质相互作用网络进行研究。研究的内容主要包含如下几个方面：

1. 蛋白质网络的研究

蛋白质相互作用网络的研究主要包括蛋白质相互作用的预测，蛋白质网络构建方法的研究以及对网络拓扑特性的研究。

虽然目前基于实验方法产生了大量的蛋白质相互作用数据，但是这些数据当中存在大量的错误包括假阳性和假阴性。通过计算方法来预测蛋白质间的相互作用，一方面可以为实验验证提供依据，节省实验预测的时间；另一方面可以作为实验方法的补充，预测实验方法没法预测的相互作用。比如实验方法很难检测到跨膜蛋白质的相互作用，而且目前的实验方法偏向于检测非瞬时性的相互作用。此外，计算方法预测的相互作用会给相互作用设定一个表示可靠性的权值。这个权值对后续的网络研究工作有很大帮助。

蛋白质相互作用预测的计算方法主要分为两大类。第一类方法是模型驱动的方法。假定蛋白质之间的相互作用服从一定的抽象模型。例如 M. Deng 等[6]假设

如果两个蛋白质存在发生相互作用的蛋白质结构域，那么这两个蛋白质之间存在相互作用。基于这样的假设就可以采用机器学习的方法来预测蛋白质相互作用。蛋白质都是由若干个蛋白质结构域组成的。首先将已知的蛋白质相互作用作为训练集，得到相互作用的结构域。对于新的蛋白质对，根据它们是否包含相互作用的结构域来确定是否存在相互作用。第二类方法是基于连带犯罪的原则，完全不用网络数据，而是基于其他生物数据来预测蛋白质相互作用。比如在相同细胞器中的蛋白质存在相互作用、拥有相同功能的蛋白质存在相互作用、基因表达谱相似的蛋白质间存在相互作用等等。这样有一系列的方法把蛋白质相互作用预测问题作为分类问题，结合各种生物特性[7]，利用贝叶斯分类器[7]，马尔可夫随机域[8]以及支持向量机[9]等方法来预测蛋白质间的相互作用。

在蛋白质网络构建方面的研究热点之一是动态蛋白质网络的构建。因为蛋白质间的相互作用并不是一成不变的，而是会随着时间、空间以及外界环境的变化而发生变化。因此蛋白质相互作用网络不仅仅是一个稳定的、静态的蛋白质和它们之间的相互作用的集合，而且还包含着随着时间和环境变化的相互作用。因此如何设计计算方法来构建动态的蛋白质相互作用网络成为人们研究的重要内容。近年来，人们开始关注动态蛋白质相互作用网络的构建，一方面是因为疾病的发展是动态的，在不同时期基因和蛋白质的表达是有差异的。构建动态蛋白质相互作用网络可以通过基因以及它们的蛋白质产物在表达上的差异来说明疾病的发展程度，从而可以提取疾病在某一段时间内的分子特征，最终有利于临床诊断以及预防性治疗。另一方面复杂疾病药物的开发也从标靶单个蛋白质或基因转移到标靶以系统为基础的动态的蛋白质子网。此外，动态蛋白质网络的构建也利于提高蛋白质复合物识别的性能和关键蛋白质识别的准确性。

目前构建动态蛋白质相互作用网络的方法主要分为两大类。一类是基于蛋白质出现的动态性。另一类是基于蛋白质共表达的变化。基于蛋白质出现的动态性的方法：利用基因表达的时间序列数据，识别在每一时刻点活性的蛋白质。然后这些活性的蛋白质节点就构成了这个时刻活性的蛋白质子网。这类构建蛋白质动态相互作用网络方法的区别主要在于识别活性蛋白质节点方法的不同。U. de Lichtenberg[10]等将蛋白质分为周期性表达的蛋白质和非周期性表达的蛋白质。非周期性表达的蛋白质组成的网络为静态的蛋白质相互作用网络。周期性表达的蛋白质将其在细胞周期的时间序列表达谱中出现峰值的时候确定为活性时刻。Hegde[11]等利用大肠杆菌在四种条件下的基因表达数据来构造动态网络。在每种条件下的表达数据中，认为信号强度高于或等于这个条件下所有节点信号强度中位值的节点是活性的。然后将这些活性节点映射到静态网络中，构建在不同条件下的动态相互作用网络。Tang 等[12]通过对酿酒酵母代谢周期表达谱的分析，采用

一个固定的阈值来识别蛋白质表达的时刻。在每一个时刻点，表达水平高于这个阈值的蛋白质被认为是表达的，从而构建这个时刻的子网。但是这个方法是基于固定阈值的。在不同表达数据上，很难选择合适的阈值。事实上，有些表达水平低的 mRNA 也会合成蛋白质，而这些都会被 Tang 的方法过滤掉。Wang 等[13]为每个蛋白质设定一个阈值。根据 3 - sigma 准则，基于每个蛋白质的基因表达数据来识别这个蛋白质的活性时刻点，从而找到各个时刻点的活性蛋白质。基于蛋白质共表达变化的动态蛋白质相互作用网络的构建方法通常是计算两个蛋白质基因表达数据之间的皮尔逊相关系数[14]。当两个蛋白质表达数据的皮尔逊相关系数高于某一个阈值时就认为它们是共表达的。将共表达的蛋白质映射到静态网络中，从而获得共表达子网。

在构建好的蛋白质相互作用网络上，对网络本身拓扑结构的研究也引起了人们的兴趣[15]。通过对酿酒酵母、果蝇等蛋白质相互作用网络的研究，发现这些网络都有着共同的特性。比如蛋白质相互作用网络中的节点服从幂律分布，也就是说大部分节点的连接度比较低，只有少部分节点的连接度比较高。这也就是所谓的无标度性。在具有无标度性的网络中移除连通度小的蛋白质对整个网络的影响不大。但是连通度大的蛋白质（Hub 节点）对外界的攻击是很脆弱的。Jeong 等[16]通过对酿酒酵母的蛋白质相互作用网络的研究，发现酿酒酵母网络能够容忍单个低度基因的敲除。但是如果移除高连通度的节点将会增加网络的直径，甚至会导致网络的瓦解。这类高连通度的基因在生物上对应的是关键基因。度中心性用来刻画蛋白质节点在网络上的重要程度。介数中心性（Betweenness Centrality）是另外一种描述网络全局拓扑特性的指标。它能够描述这个节点如何影响两个节点间的连通情况。一个节点的介数中心性定义为经过一个节点的最短路径的条数与经过这个点所有路径的比值。考虑到生物网络的模块性和层次性[17-19]，介数中心性是一个重要的刻画非 Hub 节点重要性的拓扑特征。结合 mRNA 表达数据，Han 等[18]进一步研究蛋白质相互作用网络中 Hub 节点的特征。通过计算 Hub 节点跟它们邻居节点 mRNA 表达值的关联关系，可以把 Hub 节点分为两类。一类是与邻居节点的 mRNA 表达水平密切相关的 Hub 节点，称为 Party Hub。另一类是与邻居节点的 mRNA 表达水平关联关系不是很紧密的 Hub 节点，称为 Date Hub。Party Hub 与网络中的局部功能密切相关，常被认为在功能模块内部有很强的作用。而 Date Hub 则与网络全局拓扑结构有关，担任的是全局的功能。比如酿酒酵母 Cmd1 是一个 Date Hub，连接着与恒定性阳离子（homoeostasis of cations）蛋白质折叠和稳定、发育和内质网相关的功能模块。而 Party Hub Vtil 则只是在一个内质网功能模块中出现[17]。Han 等指出，如果攻击网络中的 Date Hub，将会导致网络瓦解。而如果只是攻击 Party Hub，对网络中特征路径长度的影响与随

机地破坏网络中节点的影响差别不大。除了对网络节点的研究，He 和 Zhang[20] 还研究网络中关键边的特征。大多数网络模型认为网络中的边是同等重要的。然而区别地利用网络中的相互作用会提高人们对网络拓扑结构的理解。He 和 Zhang 将蛋白质相互作用网络中连接两个关键蛋白的边定义为关键边。通过研究分析，他们发现蛋白质的关键性不仅跟它们的连通度有关，还跟它们的连接边的关键性有关。Party Hub 的关键性很有可能与它们失去了一些关键的相互作用有关。考虑到大部分工作研究蛋白质相互作用网络没有涉及原子角度（蛋白质作用面），Kim 等[21] 利用蛋白质作用面的信息进一步分析网络中的 Party Hub 和 Data Hub 的特性。事实上，蛋白质的作用面信息可以用来研究蛋白质如何发生相互作用。比如 Aloy 和 Russell[22-24] 利用蛋白质结构域的知识来解释用双杂交实验得到的相互作用数据。Kim 的研究将 Hub 节点分为单作用面和多作用面的 Hub 节点。单作用面的 Hub 节点与邻居节点通过一个作用面发生作用，这类节点对应于 Data Hub。多作用面的 Hub 节点与多个蛋白质同时发生作用，对应于 Party Hub 节点。他们的研究表明，多作用面的 Hub 节点比其他节点进化要缓慢。它们与邻居的基因表达水平的关联程度高于单作用面节点与其邻居的基因表达水平的关联程度。

2. 基于蛋白质相互作用网络的关键蛋白质识别

关键蛋白质也叫做致死蛋白，是一类没有它们就会导致细胞死亡或者无法繁殖的蛋白质。因此生物学家致力于识别关键蛋白质，主要是出于以下两个方面的目的。从理论的角度，识别关键蛋白质有助于理解细胞生存和发展的最低需求。这样识别关键蛋白质对最近出现的合成生物学起到了至关重要的作用。因为合成生物学的目的就是创建一个具有最小基因组的细胞[25]。从实践的角度，因为关键蛋白质对于一些细菌存活必不可少，因此它们也是新的抗生素的药物标靶[26]。此外，研究结果表明，关键蛋白质（关键基因）也与人类的致病基因关系紧密[27]。研究关键蛋白质也有助于识别致病基因。在生物上，有很多实验方法来识别关键蛋白质。比如单个基因敲除[28]，RNA 干扰[29] 和有条件基因敲除[30]。然而这些实验方法既浪费时间，又效率低下，而且只能在少量物种上实行，因此迫切需要提出可靠性高的计算方法来识别关键蛋白质。

目前有一类计算方法主要是结合蛋白质相互作用网络来识别关键蛋白质。研究发现，在酵母、秀丽隐杆线虫、果蝇等物种中，那些在蛋白质相互作用网络中拥有大量的相互作用的蛋白质是更倾向于导致细胞死亡的蛋白质[16,31]。因为如果将这些节点从网络中移除，将会导致整个网络瓦解。这就是所谓的中心性致死性原则[16]。尽管人们对这一发现还存在着一些争议[20,32,33]，但是大多数研究人员还是赞同蛋白质的关键性跟它们在蛋白质相互作用网络中的拓扑特性之间存在着紧密的联系。近年来，许多利用蛋白质在蛋白质相互作用网络中的拓扑特性来

预测关键蛋白质的方法被提出来。Estrada 等[34]比较了预测蛋白质关键性的六种中心性方法［度中心性（Degree Centrality，DC）］[31]，介数中心性（Betweenness Centrality，BC）[35]，接近性中心性（Closeness Centrality，CC）[36]，子图中心性（Subgraph Centrality，SC）[37]，特征向量中心性（Eigenvector Centrality，EC）[38]和信息中心性［（Information Centrality，IC）[39]］的性能，发现这些方法在酿酒酵母蛋白质相互作用网络中的预测效果比用随机方法好很多。其中 SC 的预测效果最好。此后，Park and Kim[40]在两种酿酒酵母蛋白质相互作用网络中比较了 40 种不同的中心性算法的关键蛋白质识别性能。他们发现局部的中心性方法比全局的中心性方法识别性能更好。Lin 等[41]通过分析酿酒酵母相互作用网络中蛋白质关键性和它邻居节点的关键性，提出了 MNC（Maximum Neighborhood Component）和 DMNC（Density of Maximum Neighborhood Component）算法。Li 等[42]通过分析那些有高连通性的非关键蛋白质，发现这些非关键蛋白质的邻居节点之间几乎不存在相互作用。因此，他们提出了一个新的局部方法 LAC（Local Average Connectivity）。Wang 等[43]从关键边的特性的角度进行研究，提出了一种新的基于边聚集系数的识别关键蛋白质的中心性算法（Edge Clustering Coefficient Centrality，NC）。与以往中心性算法不同，NC 既考虑到了点的中心性，又考虑到了边的特性。尽管已提出了大量中心性方法来预测关键蛋白质，但是他们预测的结果当中只有少量的重叠，表示我们可以结合这些中心性算法去获得更好的预测效果[44]。G. del Rio 等[45]在酵母中基于 18 个不同的代谢网络，分析了 16 个不同中心性算法的关键蛋白质预测性能。实验结果表明结合至少两种中心性算法能够提高关键蛋白的预测性能。然而，结合 3 ~ 4 种中心性算法对预测性能的改变不会很大。Chua 等[46]结合了现有的几种预测方法［DC，CC，邻居功能中心性（Neighborhood Functional Centrality，NFC）和功能差异性（Functional Diversity，FD）］，提出了一种新的预测关键蛋白的概率模型。

由于蛋白质相互作用数据存在不可靠性和不完整性，会对那些基于网络拓扑特性来预测关键蛋白的方法产生影响，因此，这些中心性的方法仅仅应用于一些相互作用数据相对完整的物种中，如酿酒酵母或者大肠杆菌。其次，这些算法大多数仅仅使用了网络的拓扑特性而很少去分析已知关键蛋白质的内在的生物特性。最近一些研究人员通过结合蛋白质的拓扑特性和其他生物信息来预测关键蛋白。Tew 等[47]结合蛋白质的功能信息和拓扑信息去预测关键蛋白质。他们假设位于功能模块中心位置的蛋白质执行着关键的功能，比起其他位于外围的蛋白质，将它们移除对功能模块的影响要大得多。因此，他们基于 GO Term 相似性定义两个蛋白质的功能相似性，提出了一个新的中心性算法——邻居功能中心性。Li 等[48]结合功能相似性和逻辑回归模型，构建一个加权的蛋白质相互作用网络。

基于这个加权的蛋白质相互作用网络，他们定义了六种加权的中心性算法（DC、BC、CC、SC、EC、IC）。由于关键蛋白质更倾向于与关键蛋白质发生作用，而且常常成簇存在[49]，Ren 等[50]通过结合网络拓扑特性和蛋白质复合物信息，提出了一种新的关键蛋白质预测方法。他们使用一个蛋白质的子图中性（SC）来度量该蛋白质在蛋白质相互作用网络中的重要性。通过蛋白质在所属蛋白质复合物中的入度之和来结合它的复合物信息。此外，考虑到在同一个功能模块中蛋白质往往是共表达的，Li 等[51]提出了一种新的结合蛋白质相互作用数据和基因表达数据的中心性算法。因为关键蛋白质相比非关键蛋白质有更强的保守性，两个关键蛋白质之间更容易存在相互作用，Peng 等[52]结合蛋白质相互作用网络和蛋白质的同源信息提出了一种预测关键蛋白质的方法——ION。

3. 基于蛋白质相互作用网络的致病基因识别

遗传性疾病常常是由单个基因或多个基因发生变化引起的。超过 1800 种遗传性疾病，如镰状细胞贫血症、马凡氏综合征、亨廷顿氏病，是由单一基因的突变引起的[53]。然而，大多数这类遗传性疾病是非常罕见的。与此相反，对公众健康具有重大意义的许多疾病，如癌症、糖尿病和心血管疾病是由多个基因及其蛋白质产物与环境的相互作用引起的[54]。识别致病基因，理解基因的功能，基因间的相互作用和代谢通路在人类遗传学研究中仍然是一个重要的挑战。识别致病基因可以在临床医学上为产前产后疾病诊断提供支持，也为易感家庭早期预防和诊断治疗提供方案。

目前主要有两种产生候选致病基因的方法。第一个是连锁分析和关联研究，目的是找出候选基因与已知的致病基因在 DNA 序列标记的相对位置。第二个是全基因组关联研究（GWAS），检测单核苷酸多态性（SNP）和遗传性疾病之间的关联[55]。这两种方法都可以找到成千上万的候选疾病基因，但如何从这些候选基因中确定最有可能的致病基因，是分子生物学家和医学遗传学家面临的一个巨大挑战。由于实验方法的局限性，如需要花费很长的时间和大量的劳力，因此迫切需要有效的计算方法来解决这个问题。计算方法识别致病基因的核心思想是依据连带犯罪的原则，参照一些参考基因也叫做种子基因或训练集，采用计算方法来分析候选基因跟种子基因的关系，从中找到关系最为紧密的候选基因作为预测的致病基因。在分析关联关系时，人们会利用不同的数据资源，比如基因表达数据，基因调控数据，Pubmed 数据库中的文献资料，以及蛋白质相互作用网络的信息。

随着高通量实验技术的采用，产生了大量的蛋白质相互作用数据，越来越多的人开始研究致病基因在蛋白质相互作用网络上的特性。大量研究表明，与相同疾病关联的基因，在它们的蛋白质相互作用网络上常常是紧密连接的。Xu 等[56]

基于文献引用的蛋白质相互作用网络，对 OMIM 数据库中的致病基因进行分析后发现，OMIM 里面的基因通常具有较大的连通度，倾向于与其他疾病基因相连，并且共有邻居，彼此间的通信距离较短。因此，近年来人们利用致病基因在蛋白质相互作用网络上的特性，提出了大量识别致病基因的方法。总体而言，这些方法从拓扑特征的角度可以分为两类。一类是利用局部信息的方法。这类方法基于致病基因的直接邻居或者最短路径来推断候选的致病基因[57]。M. Oti 等[58]首先获取致病基因在蛋白质相互作用网络中的直接邻居节点，然后利用 Ensembl 数据库里面的信息获得这些邻居节点在染色体上的位置。如果它们在染色体上的位置是一个疾病基因位点时，就认为这个邻居节点是候选的致病基因。他们的研究发现有 10% 的与致病基因直接相连的邻居节点也参与同一种疾病。L. Franke 等[59]提出了一个新方法——PRIORITIZER，在预测的蛋白质相互作用网络上，多个致病基因共同邻居的基因为候选的致病基因。Li 等[60]结合蛋白质相互作用网络数据和基因表达数据，先是基于最大相关最小冗余算法（maximum relevance minimum redundancy，mRMR）分析基因表达数据，得到了 6 个候选的直肠癌致病基因。然后找到这 6 个候选基因两两间的最短距离。在这些最短路径上的 35 个基因也被列为候选的致病基因。与基于局部信息的方法不同，另外一类方法是基于网络的全局信息。这类方法模拟信息流在细胞内的流动来衡量已知的致病基因和候选基因间的连通性和亲疏关系。基于与相同疾病基因相关联的蛋白质在蛋白质相互作用网络中拥有相似的拓扑特征，Erten 等[61]提出运用 VAVIEN 方法利用蛋白质网络拓扑的相似性来推荐候选的致病基因。这项研究是将带重启的随机游走模型（random walk with restart）和皮尔逊相关系数（Pearson correlation coefficient，PCC）相结合，从全局角度评估两个蛋白质在蛋白质相互作用网络上拓扑结构的相似性。Gonçalves 等[62]则利用热核扩散模型和 PageRank 算法来得到各个候选基因全局的拓扑得分。这项研究表明全局的拓扑特征比起直接邻居和最短路径的预测效果要好。

除了在不同拓扑特性上的研究外，最近研究者关注于将蛋白质相互作用网络和其他生物信息相融合来识别致病基因。研究表明，表型相似的疾病通常会共享一些功能相似的基因[63]，因此人们构建了疾病相似性网络结合蛋白质相互作用网络信息来推荐候选的致病基因。这类方法有 Wu 等[64]基于线性回归模型在两个网络的基础上提出的 CIPHER 算法。Zhang[65]采用贝叶斯回归模型，利用疾病显性和基因亲疏关系的线性关系来推荐候选的致病基因。Yao 等[66]在 CIPHER 的基础上，利用全局信息进一步提高 CIPHER 的性能。Vanunu 等[67]提出了 PRINCE 算法，在构建的疾病相似性网络和蛋白质相互作用网络上，采用网络传播的策略来识别致病基因。此外，GO 功能注释的信息[68]以及代谢通路[69]的信息也被融

合进来推荐候选的致病基因。

4. 基于蛋白质相互作用网络的蛋白质功能预测

正确地预测蛋白质功能有助于我们从分子水平理解生命活动。随着很多物种基因组测序工作的完成，序列数据与其功能注释之间的差异日益增大。然而，从实验角度来确定蛋白质的功能不仅耗时且花费巨大。因此很多计算方法被提出来解决这个问题。计算方法预测蛋白质的功能主要是通过查找相似的已知功能的蛋白质。其中最为常见和最为可靠的方法是查找在其他物种中的同源蛋白质。目前常用的查找同源蛋白质的方法是用 BLAST[70] 和 FAST[71] 工具来比对两个蛋白质间的序列相似性。或者通过查询一些公开数据库如 Pfam[72]、ProDom[73]、SCOP[74] 等来识别两个蛋白质共同的结构域（protein domain），而不是比对整个序列。因为有研究表明蛋白质结构域是蛋白质结构和功能的基本单位。蛋白质功能的多样性是不同数量、不同类型的结构域组合的结果[75]。然而这类方法的缺点是随着大量物种测序工作的完成，产生了大量未知功能的蛋白质。它们当中很少能够在其他物种中找到功能已知的同源蛋白质。因此有一些研究开始基于蛋白质相互作用网络来给蛋白质注释功能[76]。这些基于网络的方法是基于这样一个事实，70% 到 80% 的蛋白质和与它们在网络中相关联的蛋白质至少共享一个共同的功能[77]。早期基于蛋白质相互作用网络预测蛋白质功能的方法有 Schwikowski 等[77] 提出的 NC（neighbor counting）方法。NC 将在邻居蛋白质中出现次数最多的功能作为蛋白质预测的功能。然而 NC 的缺点是忽略了不同功能在全部蛋白质中出现的背景频率。在文献［78］中，作者通过计算所考虑功能的卡方统计量来进一步提高 neighbor counting 的性能。考虑到以前的方法预测蛋白质的功能只是考虑蛋白质的直接邻居，Chua 等[79] 分析了蛋白质的直接邻居以及间接邻居的功能信息。通过赋予直接邻居（level－1）和间接邻居（level－2）不同的权重来预测蛋白质功能[80] 以及识别蛋白质复合物[81]。此外，有方法通过考虑全局网络的一致性来预测蛋白质的功能。Vazquez 等从全局角度，在保证最大数量的相关联的蛋白质共享相同的功能的前提下预测蛋白质的功能[82]。最近 Chi 等提出了一个迭代的方法，从全局一致性的角度来给蛋白质注释功能[83]。与单个的给蛋白质注释功能的方法不同，有一类基于网络的方法把蛋白质相互作用网络划分成很多功能模块[84]。在同一个模块中的蛋白质被赋予相同的功能。这类方法的依据是蛋白质功能模块是由一群共同执行特定生物功能的蛋白质组成的。最近有一些基于模块划分的蛋白质功能预测方法被提出来。这些方法的不同在于产生功能模块的方法不同[84]。

因为蛋白质相互作用网络的不可靠性，有些研究方法结合其他生物信息来给蛋白质注释功能。这些方法[85－89] 或者是融合多种网络的信息，如基因表达网络，

基因调控网络，GO 注释相似性网络等。或者是结合大量的生物特征数据，比如序列模式，同源数据，已知的功能信息，蛋白质复合物信息等。Lin 等[90]研究发现，两个蛋白质共有邻居越多，拥有相似功能的可能性就越大。Zhang 等[91]扩展了这个共有邻居的概念。认为两个蛋白质邻居的蛋白质结构域组成相似的话就会共享相似的功能。通过将蛋白质的结构域信息与蛋白质相互作用网络的信息相结合来预测蛋白质的功能。

5. 基于蛋白质相互作用网络的蛋白质复合物及功能模块识别

大多数细胞内的生物过程都是由一群蛋白质来执行。这群蛋白质在物理上彼此交互形成了蛋白质复合物。正确检测蛋白质复合物在了解大多数细胞功能的内部机制和预测未注释蛋白质的功能中起到重要作用。在生物学中，也有一些实验方法用于检测蛋白质复合物，如串联亲和纯化与质谱（TAP - ms）[92]。然而，这些实验方法既昂贵又耗时。为了克服这些局限性，许多计算方法已作为实验方法的补充被提出来预测蛋白质复合物[93]。

其中有一组研究人员致力于从蛋白质相互作用网络中挖掘蛋白质复合物。因为在同一个复合物中的蛋白质之间交互比较频繁而且共享相同的功能[84]。这样，蛋白质复合物在蛋白质相互作用网络上对应于一个稠密的子图或簇。然而，在大规模网络中寻找高度连通的子图（团）是由 NP 完成的[94]。为了解决这个问题，人们提出了许多启发式的图聚类方法来从蛋白质相互作用网络中寻找稠密的子图或者簇。这些子图或者簇中的节点之间连接紧密，而与网络中的其他节点连接稀疏。根据是否考虑网络的整体结构，这些方法可分为两种类型：全局聚类方法和局部聚类方法。

全局聚类方法通过把蛋白质相互作用网络划分成若干子图来识别蛋白质复合物。Girvan 和 New man[95]提出了 GN 算法。GN 算法通过迭代地移除具有高边介数的边来划分蛋白质相互作用网络。另一个著名的全局聚类方法是马尔可夫聚类算法（MCL）[96,97]。MCL 基于随机游走技术，通过执行两种操作，即扩充（expansion）和膨胀（inflation）将蛋白质相互作用网络划分成若干不重叠的子网络。由于随机游走技术利用了全局的网络结构，MCL 具有强大和有效的蛋白质复合物的检测性能。然而，MCL 只能产生非重叠的子图，而蛋白质复合物是高度重叠的。很多蛋白质是多功能的，涉及不同的功能模块。为了克服 MCL 方法上的缺陷，一些基于 MCL 的方法已被提出来识别重叠的蛋白质复合物[98]。

局部聚类方法检测蛋白质复合物，通常考虑的是局部邻居，而不是全局网络。基于最大团算法（CMC），Liu 等[99]通过找到网络中所有的最大团来检测蛋白质复合物。Adamcsek 等[100]开发了一个软件包，名为 CFinder，基于团渗透算法（CPM）[101]，检测大小为 k 的团，并将共享了 $k-1$ 个节点的相邻团合并。然