SHENGWU FUZA WANGLUO WAJUE GUANJIAN JISHU YANJIU

生物复杂网络挖掘关键技术研究

林志杰 著

清华大学出版社

生物复杂网络挖掘关键技术研究

林志杰 著

清华大学出版社
北京

内 容 简 介

本书从蛋白质互作网络出发，研究生物复杂网络内部存在的链路关系。全书共分为 6 章，主要内容包括蛋白质互作网络的基础知识和研究现状、利用 BenefitRank 在加权网络上进行链路预测算法、基于图模型的蛋白质复合物识别算法、基于随机游走模型的蛋白质复合物识别算法以及全书总结。本书提出的蛋白质互作网络的链路预测方法以及蛋白质复合物识别算法，将为复杂生物网络的链路预测以及蛋白质模块识别树立一个很好的范例。

本书可供相关专业的研究生、博士生做研究课题时参考使用，也可以作为在生物信息方面进行算法研究的计算机专业人员和相关专业研究人员的参考书。

图书在版编目(CIP)数据

生物复杂网络挖掘关键技术研究/林志杰著. —北京：清华大学出版社，2019

ISBN 978-7-302-51890-7

Ⅰ. ①生…　Ⅱ. ①林…　Ⅲ. ① 智能机器人—智能模拟—研究　Ⅳ. ① TP242.6

中国版本图书馆 CIP 数据核字(2018)第 291170 号

责任编辑：汤涌涛
封面设计：杨玉兰
责任校对：王明明
责任印制：董　瑾
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市龙大印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　**印　张**：9.25　**字　数**：120 千字
版　　次：2019 年 1 月第 1 版　**印　次**：2019 年 1 月第 1 次印刷
定　　价：55.00 元

产品编号：059155-01

前　言

本书是作者在研究生物信息复杂网络方面的一些重要成果的总结，具有以下特色。

第一，内容完整、条理清晰，以生物复杂网络中的蛋白质复杂网络为主要研究对象，介绍目前蛋白质相互作用网络的研究特点、现状和作用，以及目前国内外的研究方法和研究热点，对生物复杂网络研究的来龙去脉进行了比较清晰的论述，可为从事生物复杂网络研究的相关人员提供参考。

第二，阐述和提出的算法具有坚实的理论基础，算法描述以及实验验证完整清晰，可为相关研究人员提供清晰完整的研究过程和研究结果展示，为相关计算机算法研究和生物复杂网络研究方面的人员提供科学的研究方法和研究过程。

全书共分为 6 章，主要内容包括绪论、蛋白质互作网络研究进展、利用 BenefitRank 在加权网络上进行链路预测、基于图模型的蛋白质复合物识别、基于随机游走模型的蛋白质复合物识别算法、总结。其中，前两章介绍了蛋白质互作网络的基础知识和研究现状。第 3 章从蛋白质互作网络拓扑结构特性出发，分析各种蛋白质加权网络的特性，设计了有效的蛋白质相互作用预测算法。第 4 章在预测后的蛋白质互作网络的基础上研究设计了蛋白质复合物和功能模块识别算法，重点研究生物组学

数据中的蛋白质作用组学数据，针对目前较少研究的蛋白质加权网络，定义了新颖的网络链路预测模型，提出了能有效地预测蛋白质互作网络的链路预测算法；在预测后蛋白质互作网络上，根据生物学中蛋白质复合物的结构特性，定义了新的复合物模型和算法，模型有效并且能够识别出从统计意义上证明有意义的蛋白质复合物和模块，预测了一定数量的未知蛋白质的功能。第 5 章提出一个能够过滤掉蛋白质互作网络上假阴性、假阳性数据的蛋白质复合物识别算法，将会对生物学家进行蛋白质复合物的研究具有指导意义，对生物实验发挥重要作用。第 6 章是全书的总结。

本书条理清晰，层次分明，可供相关专业的研究生、博士生做研究课题时参考使用，也可以作为在生物信息方面进行算法研究的计算机专业人员和相关专业研究人员的参考书。

本书是作者在导师朱杨勇和熊赟老师的指导下完成的，并受到复旦大学计算机科学技术学院数据科学中心全体研究人员的帮助和指正，在此表示衷心的感谢！书中引用了大量国内外学者的学术文献和研究成果，在此对他们表示衷心的感谢！

由于作者水平有限，书中难免存在不足之处，希望广大读者批评指正！

林志杰

2014 年 12 月 12 日

目　录

随着人类基因测序的完成，以及解读人体基因密码的“生命之书”的正式完成，人类基因组计划(Human Genome Project，HGP)取得前所未有的进展，也标志着一个新的生物学研究时代的到来，当代生命科学研究正式步入后基因时代(Post-Genomic Era)。生物学研究由对细胞内个别基因或者蛋白质功能的局部性研究，转移到以细胞内部全部的基因、mRNA、蛋白质及代谢产物为研究对象的各种“组学”研究，逐步把分子生物学推入系统生物学时代。但是，全基因序列的序列信息不足以解释及推测细胞的各种生命现象，蛋白质才是细胞活性及功能的最终执行者，于是科学家的研究热点又回到蛋白质上[1]。

蛋白质是由多种氨基酸按特定的排列顺序通过肽键链接而成的有一定结构的高分子化合物，它是构成细胞核组织结构不可缺少的成分，是生命活动最重要的物质基础。但是细胞中的蛋白质并不是孤立存在的，而要与其他蛋白质一起进行相互作用来行使其他功能。蛋白质相互作用在生命活动中起核心作用。蛋白质-蛋白质相互作用(Protein-Protein Interaction，PPI)决定着从转录调节到酶级联反应的几乎所有生物功能，蛋白质相互作用不仅是正常生理过程(如 DNA 复制、转录、翻译、物质代谢、信号传导以及细胞周期控制)的基础，也在病理过程中起着重要的作用。可以说，几乎所有的生物过程都是通过蛋白质相互作用精确执行的(注：“相互作用”有时简称“互作”)。

目前大多数物种的蛋白质互作网络数据都不完整，因此识别出蛋白质间相互作用的完全集对于我们理解生物细胞的生理过程及功能至关重要，这也是现代基础生物学中的热点研究问题之一。目前已经开发出多种实验技术和计算方法，能够得到大规模的蛋白质相互作用数据。但也有许多不足，如 GST-pull-down 和免疫沉淀方法的通量还不足以满足蛋白质互作网络研究的需要；酵母双杂交测定 PPI 的速度最快，但是精度不够。因此，在基因组水平上预测蛋

白质相互作用，对于功能基因组研究具有十分重要的意义，也是生命科学的前沿领域。

生物有机体的细胞中的各个蛋白质并不是孤立地完成被赋予的功能，而是在特定的时间和空间内通过相互作用形成复合物，完成某些特定的生物功能。随着一系列用于蛋白质相互作用预测和评估的新方法、新技术的提出，目前可获得的蛋白质相互作用数据迅速增长。从拓扑结构上分析蛋白质互作网络的特性，并通过这些拓扑特性从大规模蛋白质互作网络中识别蛋白质复合物，对预测蛋白质功能、解释特定的生物进程具有重要意义，也成为国内外研究蛋白质互作网络的热点课题。

蛋白质互作网络的研究课题很多，其中蛋白质相互作用预测和蛋白质复合物识别是蛋白质互作网络研究中重要的内容。本章首先介绍蛋白质互作网络及其相关研究课题概况，分析本书研究问题的背景和意义，然后给出本书所阐述的主要工作内容。

1.1　蛋白质互作网络

众所周知，生物系统是一个庞大复杂的系统，该复杂系统的性质和功能不可能由单个组成元素的性质和功能完全决定，其中起主要支配作用的是组成元素之间的相互作用，以及由此而形成的集合体。随着生命科学的进一步发展，生命科学研究的重点从基因组学转移到蛋白质组学。根据生物中心法则(Genetic Central Dogma)，记录遗传信息的基因一般都要翻译成蛋白质后才能在各种生命活动中执行其功能，因此，对蛋白质的研究显得尤为重要。以蛋白质为研究对象，形成了蛋白质组学这个新的研究领域。

在各种生物分子相互作用中，最有趣味也是最具研究价值的是蛋白质之间的相互作用，目前已经有很多研究人员(如 Chen 和 Sivachenko、Uetz 和 Finley、Ito 以及 Gavin 等)针对细胞内的蛋白

质的活动行为对于生命活动的重要作用和影响展开研究。通过生物实验测得蛋白质之间的相互作用结果也变成生物学家非常感兴趣的知识资源。另外，蛋白质领域结构的研究需要大量的计算和优化算法，研究和存储蛋白质相互作用需要基于计算机的建模、管理和数据分析工具对蛋白质相互作用数据进行合理的呈现。并且对于相互作用数据的广泛关注，将大量的蛋白质相互作用表示成一个很大的网络，利用网络模型来表达蛋白质之间的作用和空间的相互关系，利用网络模型来映射蛋白质之间的生物化学生物学关系，从而通过对网络模型的研究和开发导出一些基于网络知识对蛋白质互作网络的理解和发现，来解释现实中的生命活动和生物功能，进而借助网络的一些概念、属性以及研究复杂网络的方法来理解和认识生物系统的演化和行为。我们称该网络模型为蛋白质互作网络。

蛋白质互作网络属于蛋白质组学的研究范畴，是蛋白质组学与其他大规模科学(如基因组学、生物信息学等)交叉而成的系统生物学研究的热点问题之一。随着基因测序工作的完成，检测大规模蛋白质相互作用的实验技术迅速发展，获取蛋白质相互作用数据已变得相对容易。

1.1.1 蛋白质相互作用数据库

近年来，大量的蛋白质相互作用基础数据是由酵母双杂交、串联亲和纯化、质谱分析、蛋白质芯片和噬菌体显示等高通量蛋白质组技术产生的，可获得的蛋白质相互作用数据迅速增长。这些蛋白质组技术从蛋白质组水平寻找与目标蛋白质相互作用的蛋白质，产生了大批量 PPI 数据，并已经被收集整理在不同的公共数据库中，为深入研究蛋白质网络打开方便之门。

截至目前，大概有 300 多蛋白质相互作用数据库，并且还处于增长中，成为生物学网络和通路构建的主要资源。这些蛋白质相互

作用公共数据库，也是目前使用最广泛、数据信息最完善的公共数据库。

1. DIP 数据库

DIP(Database of Interacting Proteins，蛋白相互作用数据库)专门用于存储经实验证实的来自文献报道的二元 PPI，以及来自 PDB(Protein Data Bank，蛋白质数据库)的蛋白质复合物，其目的在于建立一个简单、易用、高度可信的 PPI 公共数据库。DIP 数据库的 PPI 包含果蝇、酵母、家鼠、挪威鼠、人等多个物种，提供多种查询方式，用户可直接基于蛋白质名称、物种查询相互作用蛋白质，也可基于序列匹配的 BLAST 搜索和模体(Motif) 搜索、查询相互作用蛋白质。JDIP 是 DIP 数据库提供的一个基于 Java 语言的可视化应用工具，可把 PPI 数据以网络形式更加直观地展现出来。近年来的统计数据显示，DIP 数据库存储了通过 62 846 次实验获得的 19 200 个不同蛋白质之间的 55 889 个相互作用数据。DIP 数据库是一个关系数据库，共构建 5 个主要的数据库表，存储蛋白质、实验以及相互作用数据。DIP 数据库中存储着每个蛋白质的描述信息，如基因名称、所处细胞的位置等，还存储了具体测定蛋白质相互作用数据时，所做实验的过程信息和个人实验信息，其中存储的每一个蛋白质相互作用都有唯一编码。用户可以不同格式(如 XML，制表符分隔格式)下载 DIP 数据库的数据子集。另外，DIP 数据库可以 HUPO、PSI-MI 格式导出数据。

2. BIND 数据库

BIND(Biomolecular Interaction Network Database，生物分子互作网络数据库)是 BOND(Biomolecular Object Network Databank，生物分子对象网络数据库)的一个子数据库，收录已知的生物分子之间的相互作用，不仅包括蛋白质之间的相互作用，也包括蛋白质与 DNA、RNA、小分子、脂质以及糖类物质之间的相互作用。BIND

数据库每日更新，覆盖面广，包含人、果蝇、酵母、线虫等物种的PPI。在 BIND 数据库中，PPI 被分成三大类：二元分子相互作用(Binary Interactions)、分子复合物(Molecular Complexes)以及生物途径(Biological Pathways)，从不同层次表示分子间的相互作用关系。BIND 数据库来源于高通量技术测得的实验数据和人为的从科学文献中获取的数据，共包括 175 000 个蛋白质相互作用数据。该数据库允许不同模式的搜索，如利用生物数据库识别器、文献信息、分子结构、基因信息或基因功能等形式都可以对数据库中的数据进行检索，检索到的数据信息会通过 BIND 数据库可视化工具进行显示，非常直观，并且界面友好。

3. MINT 数据库

MINT(Molecular Interaction Database，分子相互作用数据库)\upcite{NAR:28}建立的目标是提取文献信息，存储经实验证实的生物分子相互作用。目前，MINT 数据库主要存储蛋白质物理相互作用，尤其强调哺乳动物的 PPI，同时包含部分酵母、果蝇、病毒的 PPI。目前 MINT 数据库中存储了 102 751 个相互作用，其中包括：41 768 个酵母菌蛋白相互作用，23 142 个果蝇相互作用，22 458 个哺乳动物相互作用，4 742 个线虫相互作用。在查询时，MINT 数据库可根据蛋白质名称、各数据库 ID(如 UniProtKB、PDB、Ensembl、FlyBase、OMIM)、关键词等进行基本查询，也可与 DIP 数据库一样，按照序列 BLAST 查找同源相互作用。MINT 数据库支持平面文件格式、PSI-MI 格式、Osprey 格式，并提供基于 Java 语言的网络可视化应用工具 MINT Viewer。

4. MIPS 数据库

MIPS(Mammalian Protein-Protein Interaction Database，哺乳动物的蛋白质相互作用数据库)[2]同样利用文献挖掘技术，专门存储哺乳动物的 PPI，主要包括人、大鼠、小鼠等物种。该数据库详细记

录了蛋白质相互作用的类型、实验证据及其结合位点，提供蛋白质名称、实验方法、物种等多种查询方式，查询结果以短格式(Short Format)和长格式(Long Format)两种形式显示。前者以列表形式显示相互作用蛋白质，后者则详细描述 PPI 信息，包括 PPI 的参考文献、实验证据、结合位点、生物学功能等。MIPS 也提供 PPI 可视化网络分析，同时可链接到由同一研究小组开发的小鼠基因组数据库 PEDANT。

5. IntAct 数据库

IntAct 也是一个存储和分析生物分子间相互作用的公共数据库[3]，主要记录二元相互作用及其实验方法、实验条件和相互作用结构域，包括人、酵母、果蝇、大肠杆菌等物种。IntAct 数据库分为基本查询和高级查询：基本查询可以根据蛋白质名称、PubMedID 等进行简单搜索；高级查询根据实验方法和 IntAct 自定义的控制词汇(Controlled Vocabularies)进行查询，查询结果更加精确。IntAct 数据库支持 PSI-MI XML 1.0 和 PSI-MI XML 2.5 的格式，提供 PPI 网络的可视化在线分析，同时支持 Cytoscape、Proviz 等第三方网络构建软件。除了存储并查询相互作用蛋白质信息，IntAct 数据库可基于“Pay As You Go”算法预测下拉(Pull-Down)实验的最佳诱饵蛋白。IntAct 研究小组还建议生物学家在文献发表之前向该数据库直接提交 PPI 信息。这一过程如同向 GeneBank 数据库直接提交核苷酸序列一样，极大地方便了数据的增加和管理。

1.1.2　算法预测的蛋白质相互作用数据库

高通量的生物实验技术正在以不断增长的速度检测蛋白质之间的相互作用，不断丰富蛋白质互作网络，生物学家也不断对蛋白质互作网络做出新的阐释。然而在实验过程中，真核生物的复杂性已经开始阻碍高通量实验技术的顺利进行。

新的研究开始证实现有蛋白质相互作用数据库的正确性和可信性。目前已设计了很多算法来预测蛋白质之间的相互作用，这些算法都是以证实的相互作用作为算法的输入，产生一些基于生物信息考量的相互作用。这些方法基于不同的生物考量，但是采用的方式和方法基本一致。比如算法输入一个相对低级的真核生物 A 和 B，发现具有相近甚至相同功能的其他物种(如人类)的 A 和 B，然后基于功能信息评估它们之间的相互作用。

例如，OPHID 数据库是映射基于模型组织的实验相互作用数据到人类相互作用数据获得的相互作用数据。类似的 POINT 数据库是验证与人类直系同源的相互作用数据，然后基于功能信息过滤得到的相互作用数据信息。相反的，IntNetDB 基于复杂的概率模型结合不同的生物信息——mRNA、共表达和序列相似性，预测得到的相互作用数据。

1. OPHID 数据库

OPHID(Online Predicted Human Interaction Database，在线预测人类相互作用数据库)主要包括人类蛋白质相互作用数据，这些人类蛋白质相互作用信息是从文献以及 MINT、BIND 数据库中获取数据，然后对 Saccharomyces Cerevisiae(啤酒酵母)、Caenorhabditis Elegans(秀丽隐杆线虫)、Drosophila Melanogaster(黑腹果蝇)及 Mus musculus(小家鼠)等组织进行蛋白质相互作用预测得到的预测相互作用数据。OPHID 数据库基于蛋白质相互作用复合进化论的假设，从进化论方面上升到理论的可能性角度，我们可以把从实验测定的相互作用模型映射到人类蛋白质相互作用模型上，预测人类蛋白质之间的相互作用。不过这种映射的合理性还只是停留在方法论的层面。OPHID 数据库构建分为两个步骤：第一步，利用 BLASTP[32] 确立两个蛋白质相互作用之间是否存在直系同源交互关系；第二步，如果存在正交关系，则建立人类蛋白质相互作用映射。为进一步提高蛋白质之间的相互作用存在的可能性，预测相互作用时还考

虑到 3 个参数：蛋白质领域、基因共表达、基因本体(Gene Ontology，GO)生物语义词典。根据最近的统计，OPHID 数据库中包含 10 682 个蛋白质，以及 49 008 个这些蛋白质之间的相互作用。

OPHID 数据库中的蛋白质以及蛋白质相互作用可以通过蛋白质 ID 进行查询，其相互作用结果可以利用图可视化软件进行可视呈现。该数据库对以学术研究为目的的用户免费开放。查询 OPHID 数据库和可视化相互作用网络的软件工具 NAViGaTOR(Network Analysis, Visualization, Graphing TORonto)可以免费下载。并且 OPHID 数据库支持平面文件格式、PSI-MI 格式数据的导入和导出。

2. POINT 数据库

POINT(Prediction of Interactome Database，相互作用组预测数据库)[33]存储的也是人类蛋白质相互作用数据。该数据库中的蛋白质相互作用数据集是从直系同源交互数据集中得到的。该数据库利用 Worm、Fly 和 Yeast 为预测的起始点，然后投影到人类蛋白质数据点上，通过拓扑结构(空间共定位分布)、时间和基因本体的功能注释信息，对相互作用对的模式进行进一步优化[34]。该数据库能够访问 Mouse、Fruit、Fly、Worm 和 Yeast 数据库。

POINT 数据库中的蛋白质相互作用能以容易被人理解的图模型的方式展示，或者以树形结构呈现，同时 POINT 数据库对蛋白质相互作用添加了一些空间接近度、时间同步性和 GO 注释的信息。

3. IntNetDB 数据库

IntNetDB(Integrated Network Database，集成网络数据库)是整合的网络数据库，通过概率模型整合各种类型的功能数据，来预测蛋白质相互作用。该数据库整合了 27 个基因组、蛋白质组和功能注释数据集的信息，来预测人类蛋白质互作网络。目前该数据库中存储了 9901 个蛋白的 180 000 个相互作用数据。但是，IntNetDB 中不包含预测的 Yeast 蛋白质相互作用，不过可以参考其他数据库

中 Yeast 蛋白质相互作用数据，这要由使用该数据库的作者来选择。为了学术界对生物信息能进行深入的研究和分析，IntNetDB 数据库也提供了一个软件工具，用于从一些指定的网络上抽取拓扑结构上高度链接的网络数据。

4. STRING 数据库

STRING(Search Tool for the Retrieval of Interacting Genes/Proteins，相互作用基因、蛋白质查询检索工具)数据库是大融库，其目的是将蛋白质-蛋白质、蛋白质-DNA，以及 DNA-DNA 各种生物相互作用都存储在一个大集合库中。在 STRING 数据库中，各种生物关联都建立在分子之间的物理相互作用和间接作用基础上。例如两个蛋白质参与了同一个通路。STRING 数据库将从其他数据导入的作用数据与 Denovo 预测的关系合并在一起，其中预测的作用关系是基于功能预测，基于功能的预测主要是基于基因邻居和基因融合事件，还有基因组数据中的基因同现来预测相互作用并进行存储。

STRING 数据库可以通过 Web 网站专门的蛋白质识别工具或者输入蛋白质序列来访问，如果该蛋白质存在于数据库中，一些有关该蛋白质的相互作用关系会呈现在一个可视化的所谓的预测主窗口上，而且展示的关系图利用不同颜色的边区分不同的关系类型。用户可以浏览窗口上显示的结果，也可以继续访问有关蛋白质相互作用的一些实验证据，还可以将查询到的蛋白质相互作用数据以平面文件的形式下载到本地，甚至可以下载到整个数据库。

总之，蛋白质相互作用数据管理与其他领域的相互作用数据管理具有相似的问题，即 PPI 相互作用数据需要存储、转换、查询和分析。另一方面，PPI 数据利用图模型展示，又提出了一些新的问题。下面主要讨论蛋白质相互作用数据库的存储管理问题。很多科学家在蛋白质相互作用数据存储的 HUPO PSI-MI 标准格式上投入了大量精力，但是目前大多数的蛋白质相互作用数据以二元相互作用存储，没有考虑基于 XML 语言和相关的 XML 数据库。