同济大学 1907-2017
Tongji University

同济博士论丛
TONGJI Dissertation Series

总主编 伍 江 副总主编 雷星晖

邓 磊 关佶红 著

基于机器学习的蛋白质相互作用与功能预测

Protein Interaction and Function Prediction Based on Machine Learning Techniques

同济大学出版社
TONGJI UNIVERSITY PRESS

同济博士论丛
TONGJI Dissertation Series
总主编 伍 江 副总主编 雷星晖

邓 磊 关佶红 著

基于机器学习的蛋白质相互作用与功能预测

Protein Interaction and Function Prediction Based on Machine Learning Techniques

同济大学出版社
TONGJI UNIVERSITY PRESS

内 容 提 要

本书采用机器学习的方法，研究了蛋白质相互作用和功能预测的几个重要方面：蛋白质相互作用位点预测、蛋白质相互作用能量热点（HotSpots）预测、蛋白质相互作用预测和蛋白质功能预测，提出一系列的蛋白质相互作用及功能预测方法。

本书适合相关专业的高校师生阅读使用。

图书在版编目（CIP）数据

基于机器学习的蛋白质相互作用与功能预测/邓磊，关佶红著. — 上海：同济大学出版社，2018.5

（同济博士论丛/伍江总主编）

ISBN 978-7-5608-7831-7

Ⅰ.①基… Ⅱ.①邓… ②关… Ⅲ.①蛋白质-研究 Ⅳ.①Q51

中国版本图书馆 CIP 数据核字（2018）第 088199 号

基于机器学习的蛋白质相互作用与功能预测

邓 磊 关佶红 著

出 品 人 华春荣 责任编辑 王有文 熊磊丽

责任校对 谢卫奋 封面设计 陈益平

出版发行 同济大学出版社 www.tongjipress.com.cn

（地址：上海市四平路 1239 号 邮编：200092 电话：021-65985622）

经 销 全国各地新华书店

排版制作 南京展望文化发展有限公司

印 刷 浙江广育爱多印务有限公司

开 本 787 mm×1092 mm 1/16

印 张 11.5

字 数 230 000

版 次 2018 年 5 月第 1 版 2018 年 5 月第 1 次印刷

书 号 ISBN 978-7-5608-7831-7

定 价 74.00 元

“同济博士论丛”编写领导小组

组　　长： 杨贤金　钟志华

副 组 长： 伍　江　江　波

成　　员： 方守恩　蔡达峰　马锦明　姜富明　吴志强
徐建平　吕培明　顾祥林　雷星晖

办公室成员： 李　兰　华春荣　段存广　姚建中

“同济博士论丛”编辑委员会

袁万城　莫天伟　夏四清　顾　明　顾祥林　钱梦騄
徐　政　徐　鉴　徐立鸿　徐亚伟　凌建明　高乃云
郭忠印　唐子来　阎耀保　黄一如　黄宏伟　黄茂松
戚正武　彭正龙　葛耀君　董德存　蒋昌俊　韩传峰
童小华　曾国荪　楼梦麟　路秉杰　蔡永洁　蔡克峰
薛　雷　霍佳震

秘书组成员：谢永生　赵泽毓　熊磊丽　胡晗欣　卢元姗　蒋卓文

总 序

在同济大学110周年华诞之际，喜闻“同济博士论丛”将正式出版发行，倍感欣慰。记得在100周年校庆时，我曾以《百年同济，大学对社会的承诺》为题作了演讲，如今看到付梓的“同济博士论丛”，我想这就是大学对社会承诺的一种体现。这110部学术著作不仅包含了同济大学近10年100多位优秀博士研究生的学术科研成果，也展现了同济大学围绕国家战略开展学科建设、发展自我特色，向建设世界一流大学的目标迈出的坚实步伐。

坐落于东海之滨的同济大学，历经110年历史风云，承古续今、汇聚东西，秉持“与祖国同行、以科教济世”的理念，发扬自强不息、追求卓越的精神，在复兴中华的征程中同舟共济、砥砺前行，谱写了一幅幅辉煌壮美的篇章。创校至今，同济大学培养了数十万工作在祖国各条战线上的人才，包括人们常提到的贝时璋、李国豪、裘法祖、吴孟超等一批著名教授。正是这些专家学者培养了一代又一代的博士研究生，薪火相传，将同济大学的科学研究和学科建设一步步推向高峰。

大学有其社会责任，她的社会责任就是融入国家的创新体系之中，成为国家创新战略的实践者。党的十八大以来，以习近平同志为核心的党中央高度重视科技创新，对实施创新驱动发展战略作出一系列重大决策部署。党的十八届五中全会把创新发展作为五大发展理念之首，强调创新是引领发展的第一动力，要求充分发挥科技创新在全面创新中的引领作用。要把创新驱动发展作为国家的优先战略，以科技创新为核心带动全面创新，以体制机制改

革激发创新活力，以高效率的创新体系支撑高水平的创新型国家建设。作为人才培养和科技创新的重要平台，大学是国家创新体系的重要组成部分。同济大学理当围绕国家战略目标的实现，作出更大的贡献。

大学的根本任务是培养人才，同济大学走出了一条特色鲜明的道路。无论是本科教育、研究生教育，还是这些年摸索总结出的导师制、人才培养特区，“卓越人才培养”的做法取得了很好的成绩。聚焦创新驱动转型发展战略，同济大学推进科研管理体系改革和重大科研基地平台建设。以贯穿人才培养全过程的一流创新创业教育助力创新驱动发展战略，实现创新创业教育的全覆盖，培养具有一流创新力、组织力和行动力的卓越人才。“同济博士论丛”的出版不仅是对同济大学人才培养成果的集中展示，更将进一步推动同济大学围绕国家战略开展学科建设、发展自我特色、明确大学定位、培养创新人才。

面对新形势、新任务、新挑战，我们必须增强忧患意识，扎根中国大地，朝着建设世界一流大学的目标，深化改革，勠力前行！

万　钢

2017 年 5 月

论丛前言

承古续今，汇聚东西，百年同济秉持“与祖国同行、以科教济世”的理念，注重人才培养、科学研究、社会服务、文化传承创新和国际合作交流，自强不息，追求卓越。特别是近20年来，同济大学坚持把论文写在祖国的大地上，各学科都培养了一大批博士优秀人才，发表了数以千计的学术研究论文。这些论文不但反映了同济大学培养人才能力和学术研究的水平，而且也促进了学科的发展和国家的建设。多年来，我一直希望能有机会将我们同济大学的优秀博士论文集中整理，分类出版，让更多的读者获得分享。值此同济大学110周年校庆之际，在学校的支持下，“同济博士论丛”得以顺利出版。

“同济博士论丛”的出版组织工作启动于2016年9月，计划在同济大学110周年校庆之际出版110部同济大学的优秀博士论文。我们在数千篇博士论文中，聚焦于2005—2016年十多年间的优秀博士学位论文430余篇，经各院系征询，导师和博士积极响应并同意，遴选出近170篇，涵盖了同济的大部分学科：土木工程、城乡规划学(含建筑、风景园林)、海洋科学、交通运输工程、车辆工程、环境科学与工程、数学、材料工程、测绘科学与工程、机械工程、计算机科学与技术、医学、工程管理、哲学等。作为“同济博士论丛”出版工程的开端，在校庆之际首批集中出版110余部，其余也将陆续出版。

博士学位论文是反映博士研究生培养质量的重要方面。同济大学一直将立德树人作为根本任务，把培养高素质人才摆在首位，认真探索全面提高博士研究生质量的有效途径和机制。因此，“同济博士论丛”的出版集中展示同济大

学博士研究生培养与科研成果，体现对同济大学学术文化的传承。

“同济博士论丛”作为重要的科研文献资源，系统、全面、具体地反映了同济大学各学科专业前沿领域的科研成果和发展状况。它的出版是扩大传播同济科研成果和学术影响力的重要途径。博士论文的研究对象中不少是“国家自然科学基金”等科研基金资助的项目，具有明确的创新性和学术性，具有极高的学术价值，对我国的经济、文化、社会发展具有一定的理论和实践指导意义。

“同济博士论丛”的出版，将会调动同济广大科研人员的积极性，促进多学科学术交流、加速人才的发掘和人才的成长，有助于提高同济在国内外的竞争力，为实现同济大学扎根中国大地，建设世界一流大学的目标愿景做好基础性工作。

虽然同济已经发展成为一所特色鲜明、具有国际影响力的综合性、研究型大学，但与世界一流大学之间仍然存在着一定差距。“同济博士论丛”所反映的学术水平需要不断提高，同时在很短的时间内编辑出版 110 余部著作，必然存在一些不足之处，恳请广大学者，特别是有关专家提出批评，为提高同济人才培养质量和同济的学科建设提供宝贵意见。

最后感谢研究生院、出版社以及各院系的协作与支持。希望“同济博士论丛”能持续出版，并借助新媒体以电子书、知识库等多种方式呈现，以期成为展现同济学术成果、服务社会的一个可持续的出版品牌。为继续扎根中国大地，培育卓越英才，建设世界一流大学服务。

伍　江

2017 年 5 月

前 言

蛋白质是一切生命的物质基础，是细胞和机体的重要组成部分。蛋白质间的相互作用支撑和影响着生命体内各种功能的实现。研究蛋白质相互作用和功能对于理解生命活动的内在机理、疾病治疗、新药开发和蛋白质设计都具有重要的意义。随着以高通量测序为代表的分子生物学技术的飞速发展，越来越多的基因组被测序，蛋白质的序列和结构数据也快速增长，使用传统实验方法来识别蛋白质相互作用、标注蛋白质功能已远远不能满足当前的需求，因此，探索基于计算的蛋白质相互作用和功能预测新技术，并揭示其中的生物学规律已成为日益重要的研究课题。

本书采用机器学习的方法，研究了蛋白质相互作用和功能预测的几个重要方面：蛋白质相互作用位点预测、蛋白质相互作用能量热点(Hot Spots)预测、蛋白质相互作用预测和蛋白质功能预测。提出了一系列的蛋白质相互作用及功能预测方法。本书的主要研究如下：

1. 在蛋白质相互作用方面，研究了基于集成学习和基于结构邻居模板的蛋白质相互作用位点预测方法，提出了基于半监督学习和结构邻居属性的蛋白质能量热点预测方法，研究了基于结构的全基因组蛋白质

相互作用预测算法。

(1) 提出了一种有效的相互作用位点预测集成学习方法。该方法结合 bootstrap 重采用技术、基于 SVM 的融合分类器及加权投票策略，来克服样本的不平衡问题，并有效地利用了一系列的序列、结构特征。为了提高所提出方法的实用性，分别设计了两个特殊的分类器来处理缺少同源蛋白和结构信息的情况。方法的鲁棒性也通过从蛋白质表面残基和蛋白质所有残基两种情况下有效预测相互作用位点得到了验证。

(2) 由于蛋白质相互作用界面的保守性在结构邻居之间非常显著，本书提出了一种基于结构邻居模板和支持向量机的相互作用界面预测方法。通过将查询蛋白的结构邻居的已知界面残基，映射到查询蛋白的表面残基上，得到相互作用接触频率表，然后使用支持向量机来预测每个表面残基成为界面残基的分数。该方法在 DKBM 和 CAPRI 两个数据集上都比其他已有方法具有更高的性能。研究开发了相应的具有良好交互性和易用性的蛋白质相互作用界面预测 Web 服务器——PredUs。

(3) 由于丙氨酸扫描突变实验昂贵而且费时，能量热点实验数据非常少。针对这一问题，提出了一种迭代半监督能量热点预测方法——SemiHS。在少量有标记样本的基础上，通过迭代加入大量无标记样本来训练预测准确度更高的模型，并有效克服样本的不平衡问题。

(4) 开发了一种基于结构邻居特征的能量热点集成预测方法。在 108 个基于序列、结构和能量的残基特征基础上，分别计算了 108 个欧式邻居特征和 108 个 Voronoi 邻居特征，并使用随机森林的方法选择出了前 46 个重要特征。由于能量热点预测中存在不平衡问题，通过多次对负样本(非能量热点)进行采样来构建集成分类器，取得了非常好的预

测性能。在此基础上，还开发了能量热点预测 Web 服务器——PredHS。

(5) 研究了基于结构的全基因组蛋白质相互作用预测算法。首先，对于查询蛋白质对的结构或者同源模型，使用结构比对算法分别搜索出它们的结构邻居，然后对结构邻居的复合物模板进行叠加，形成相互作用模型，再使用贝叶斯网络对相互作用模型进行评估。最后，使用朴素贝叶斯方法对结构信息和其他非结构信息进行集成，建立了蛋白质相互作用综合预测模型。无论是在数据集还是在全基因组上，我们的方法比已有非结构预测方法和高通量实验方法具有更好的准确性和有效性，应用前景广阔。

2. 在蛋白质功能预测方面，分别提出了基于序列组成信息和基于结构比对及多数据源的蛋白质功能预测方法。

(1) 分别研究了四种蛋白质序列基本组成模块：N-grams、二进制谱、Pfam Domain 和 InterPro Domain。根据蛋白质序列中四种组成模块出现的频率，蛋白质序列被转化成了固定长度的高维向量，并使用支持向量机来预测基于 Gene Ontology 的功能。使用了潜在语义分析(LSA)和非负矩阵分解(NMF)来去除噪声并提高预测效率。实验结果表明，蛋白质序列组成信息可以有效预测蛋白质功能。

(2) 由于相似的蛋白质结构意味着相似的蛋白质功能，提出了一种基于结构比对的蛋白质功能预测模型——PredGO。对于查询蛋白的结构或者同源模型，首先使用结构比对方法搜索出其第一级的结构邻居，然后对于查询蛋白的序列同源，在结构邻居数据库中查询出第二级结构邻居。设计了一个有效的打分函数来对两级结构邻居的功能标注进行评估，并将分数较高的功能标记到查询蛋白上。此外，PredGO 还使用贝叶斯方法集成了蛋白质序列和相互作

用等非结构信息。实验表明，以上预测方法比已有的非结构方法具有更好的预测准确率和覆盖度，能应用到对未知蛋白质序列和结构的功能识别中。

目 录