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

王 晓 著

多标记生物数据建模与预测方法的研究

Research on Modeling and Prediction Techniques for Multi-Label Biological Data

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

王 晓 著

多标记生物数据建模与预测方法的研究

Research on Modeling and Prediction Techniques for Multi-Label Biological Data

同济大学出版社
TONGJI UNIVERSITY PRESS

内 容 提 要

本书主要围绕多标记生物数据的属性识别方法展开深入的研究。多标记生物数据的属性识别，是生物信息学中近年来新出现的一个研究领域。由于后基因组时代生物数据的爆炸式增长和它们的多标记特征，迫切需要开发出新的计算预测方法以便及时、可靠地预测出它们的多种功能或属性。

本书适于生物领域和控制/电信专业研究人员和工作人员阅读。

图书在版编目(CIP)数据

多标记生物数据建模与预测方法研究 / 王晓著. —
上海：同济大学出版社，2017.8
（同济博士论丛 / 伍江总主编）
ISBN 978-7-5608-6860-8

Ⅰ. ①多… Ⅱ. ①王… Ⅲ. ①生物信息论—数据模型—方法研究 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字(2017)第 070280 号

多标记生物数据建模与预测方法研究

王 晓 著

出 品 人 华春荣 责任编辑 吕 炜 卢元姗
责任校对 徐春莲 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 8
字 数 160 000
版 次 2017 年 8 月第 1 版 2017 年 8 月第 1 次印刷
书 号 ISBN 978-7-5608-6860-8

定 价 42.00 元

“同济博士论丛”编写领导小组

“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

多标记生物数据的属性识别，是生物信息学中近年来新出现的一个研究领域。大量研究发现，许多生物分子都拥有不止一种功能或特性，因而需要多个标记来注释其相应的属性。由于后基因组时代生物数据的爆炸式增长和它们的多标记特性，迫切需要开发出新的计算预测方法以便及时、可靠地预测出它们的多种功能或属性。本书围绕多标记生物数据的属性识别方法展开深入的研究，完成以下五方面工作：

(1) 本书把多标记学习技术引入蛋白质亚细胞定位领域，最早将蛋白质多亚细胞位置预测形式化为一个多标记分类任务，并且介绍了四种流行的能够准确反映多位置预测性能的评价指标，比较了两类多标记学习方法的性能优劣。实验结果表明，利用标记间相互关系的方法取得了比利用标记相关特征的方法更好的性能，为进一步研究奠定了基础。同时为真核与病毒两个生物体分别构造了各自专用的多位置蛋白质预测器，并提供了在线预测服务网站。

(2) 新合成或新发现的蛋白质的结构和功能尚不清楚，准确地了解它们的亚细胞位置的信息显得特别重要。针对已有方法没有考虑标记间关系，本书提出一种新颖的基于随机标记选择的预测方法 RALS；同

时为了解决新发现或合成的蛋白质无法表示成 GO 特征进而使预测性能大打折扣的问题，本书采用融合伪氨基酸组成和序列进化信息的方法来提取蛋白质的特征。实验结果表明，通过借助集成学习的思想间接地利用亚细胞位置间的相互关系，显著地提高了预测性能，并优于当时已有的最好结果。

(3) 以往研究人员主要专注于在细胞级别预测蛋白质的位置，本书更进一步研究叶绿体细胞器的亚结构，构建了一个包含多亚叶绿体位置的蛋白质数据集，提出了一种结合标记相关特征和标记间关系的预测方法。实验结果表明，通过选取与每个位置最相关的特征，并且加入了不同位置之间的相互关系，该方法能够很好地对蛋白质的多位置特性建模，因而取得更优越的性能。本研究是该领域的第一个对多亚叶绿体位置进行建模和预测的工作，为蛋白质亚-亚细胞位置预测研究提供了重要的参考价值。

(4) 大量的抗微生物肽不止有一个功能，可能同时拥有多种功能。同时识别出它们的多种功能类型，对抗生素替代药物的研制具有极其重要的意义。目前的工作大多都局限于仅能识别抗微生物肽，不能进行更深一层的多功能类型预测。本书把集成学习和多标记学习结合起来，创新地提出一种最优多标记集成分类算法来预测抗微生物肽的多种功能类型，实验结果表明，通过分别为每个标记(抗微生物肽的功能)选择不同的最优分类器组合，去除无关和冗余的分类器，显著地改进了预测性能。

(5) 为了更好地为生物学家提供服务，本书的所有研究成果都已开发成在线生物信息服务网站，使生物学家仅通过互联网和浏览器就可以方便快速地获得所需分析结果，并且为进一步指导实验设计提供强有力的支撑。同时，在线生物信息服务网站的建立，也为生物信息学家之间公开透明地进行预测算法的性能比较提供便利，可以进一步促进生物信息学的发展。

目 录

第1章 绪 论

1.1 多标记数据建模与预测概述

单标记数据集中的每个实例都由描述其概念的一个标记标注，也就是说，人为假设真实世界的对象与其概念标记是一一对应的关系。传统的分类算法，比如，支持向量机、kNN、决策树等，通过对这种仅有一个概念标记的单标记数据集进行学习，以尽可能正确地预测出训练集以外的实例的唯一概念标记。然而，真实世界的对象并不都只有一个概念标记，比如，一篇文章可以包含多个主题，一个蛋白质序列可能位于多个亚细胞位置，一个抗微生物肽可能有多种功能类型，等等。实际上，这种情况在真实世界中比比皆是。相对于单标记数据集，多标记数据集就是由同时包含多个概念标记的样本组成。由于唯一概念标记假设，传统的分类算法不再能处理多标记数据集。因此，多标记数据建模和预测方法应运而生。

过去十几年，多标记学习吸引了大量研究人员的关注，产生了大量的学习算法。目前大约有 140 篇多标记学习研究论文，其中 60 多篇发表在 2007—2012 年机器学习相关顶级会议[1,2]。研究者在 2009 年、2010 年和 2011 年先后举办了三次多标记学习的国际学术研讨会，推动其发展，多标