学校编码：10384　　分类号_____密级_____

学号：23020111153037　　UDC_____

厦门大学

硕士学位论文

基于集成学习与多标记学习的蛋白质分类方法研究

The Research of Protein Classifications based on Ensemble Learning and multi-label Learning

陈伟程

指导教师姓名：邹权 副教授

专业名称：计算机应用技术

论文提交日期：2014 年 月

论文答辩日期：2014 年 月

学位授予日期：2014 年 月

答辩委员会主席：_____

评阅人：_____

2014年 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：陈伟雄

2014 年 5 月 20 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　年　月　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：陈伟程

2014年5月20日

摘 要

随着大量生物学数据的增长，仅仅利用传统的生物学实验来测定蛋白质结构或其他性质的方法不仅需要耗费大量人力物力，其时间的耗费也是相当大。因此，如何建立自动学习的方法来预测蛋白质，从而大大降低生物实验的成本是一个非常有意义的研究课题。另一方面，随着机器学习技术的快速发展，其应用领域也得到了不断的扩宽，特别是在生物学领域的应用，面对大量的生物信息机器学习得到了充分的应用。

本文的主要内容包括以下几个方面：

（1）引入了两种蛋白质特征提取方法。本文在蛋白质分类问题当中引入了两种有效的特征提取方法：一种为代表氨基酸组成成分即物理化学特性的混合特征，共计 188 维；另一种为代表蛋白质同源信息的基于位置特异性得分矩阵的特征，共计 20 维。两种特征提取方法各有优劣：188 维特征提取速度快，但会牺牲一定的准确率；20 维特征提取较为耗时，但却具有更高的预测精度。

（2）提出了基于集成分类器的蛋白质折叠模式分类方法。蛋白质结构预测是生物信息学当中的重要课题，而蛋白质折叠模式的识别对蛋白质结构预测起到了关键作用。本文针对以往的蛋白质折叠模式分类准确率不高的情况，引入了集成分类器。本文的集成分类器基于投票的机制，最终预测结果集成了两种基分类器的分类结果，在公用数据集中取得了最佳的分类效果。

（3）提出了基于多标签学习的双层酶分类模型。第一层模型预测给定的蛋白质序列是否是酶，若是酶，第二层则进一步预测酶的功能子类。多功能酶由于其特殊的性质成为了酶分类问题当中非常棘手的异类。本文利用了机器学习中的多标签分类，解决了以往研究者未曾涉足的多功能酶分类问题，并且取得了很好的分类效果。

（4）开发了蛋白质折叠模式在线预测平台 PPL 以及酶在线预测平台 IME。PPL 和 IME 都具有在线预测功能，此外还提供程序的下载以便进行本地实验。网站中还提供本文所用到的数据集下载，方便用户使用与研究。

关键词：蛋白质分类；酶分类；多标记学习

Abstract

With the growth of overwhelming amount of biological data, using traditional biological experiments alone to determine protein structures and other properties not only requires a lot of manpower and resources, but also costs a lot of time. How to build “in silicon” methods for predicting proteins thus reduce the costs of biological experiments can be a meaningful topic. On the other hand, the rapid development of machine learning technology allows its application fields being constantly widening, especially the field of biology -- machine learning can be fully applied when facing those numerous biological information.

The main contents of this paper include the following aspects:

(1) Introducing two methods for proteins feature extraction. In this paper, we apply two effective ways to extract features for the problem of protein classification: the first is a mixed feature combining amino acid compositions and physicochemical properties, totaling 188 dimensions; while the second is based on position-specific scoring matrices (PSSM), which illustrates protein homologous information, totaling 20 dimensions. The two feature extraction methods have their own pros and cons: the former one extracts faster at the expense of lower accuracy, while the latter costs longer time to get higher prediction accuracy.

(2) Proposing a method for protein fold classification based on ensemble classifiers. Protein structure prediction is an important topic in bioinformatics, and protein fold identification plays a key role in predicting protein structures. In this paper, aiming to alter the fact that accuracy of previous models for protein fold classification is quite low, we introduce ensemble classifiers. Our ensemble classifier is based on voting mechanism and its final result acquires a best accuracy on the common data set through integrating outcomes obtaining by the two basic classifiers.

(3) Proposing a 2-layer enzyme classification model based on multi-label learning. The first layer gives answer to whether the protein is an enzyme or not, while the second further predicts functions of the enzyme. Multifunction enzyme has

become a very tricky heterogeneous because of its special properties when facing the problem of enzyme classification. In this paper, we apply the multi-label classification skill which belongs to machine learning, solving the multifunction enzyme classification problem which previous researchers have not got involved, and achieving good classification results.

(4) Developing an online prediction platform for predicting protein fold called PPL, and another for predicting enzymes called IME. PPL and IME both provide programs for local experiments, in addition to their basic function of online prediction. Data sets are also included for downloading so that users can get easily access to our data and do further research.

Key Words: protein classification; enzyme classification; multi-label learning

目录

CONTENTS

第一章 绪论

1.1 引言

1.1.1 蛋白质结构

蛋白质的结构跟蛋白质的功能息息相关，是生物信息学领域热门的研究领域。由于蛋白质是生命的物质基础，是生物机体组织、器官的重要组成部分，在生命体中占据着重要作用，通过研究蛋白质结构特点分析其生理功能对学习酶代谢、开发新药物及抗 HIV 病毒等工作都具有突破性作用。蛋白质是多种氨基酸组合而成的，蛋白质的一级结构即氨基酸序列会根据各氨基酸残基的极性、电荷表面张力、亲疏水等等多种特性通过残基间的相互作用而折叠成立体的三级结构。蛋白质折叠模式的识别是指通过氨基酸序列的一级结构即氨基酸序列来直接预测生成的三级结构，它可以充分了解蛋白质的结构功能和工作过程。但是这种短时间内完成的折叠过程却是难以十分准确地计算得出的，因此蛋白质折叠问题被列为“21 世纪的生物物理学”的重要课题，也是分子生物学还未完全解决的重大生物学问题。

随着人类基因组等项目的快速开展，越来越多的生物数据被发现，成千上万的蛋白质序列被陆续测出。然而，一条蛋白质序列需要折叠成一定的空间结构，其特定的生理功能才能得以发挥，在完成蛋白质序列测序工作以后，人们更希望能够得到这些蛋白质的空间结构，从而发现蛋白质结构与功能之间的联系。因此，蛋白质结构与功能的研究就成为了后基因组时代生命科学领域研究者们的主要研究任务和目的。

针对一条已知蛋白质序列的功能预测有两种途径：序列-序列（sequence-sequence）比对和序列-结构（sequence-structure）比对，即分别依据序列的排列方式和表现出的结构特征对功能进行分析。然而，当两条序列的序列相似度不明显时，它们可能反映了高度重叠的结构特征，此时，只有序列-结构的比对方法能够得出有效结论，因此从蛋白质的结构入手来分析其特征和功能是当前研究者所采用的普遍做法。

蛋白质结构的不同会影响到具体的生物功能，因此蛋白质结构的多样性造

就了不同的生物学功能。若想要通过蛋白质一级序列来预测蛋白类型或预测蛋白质的二级结构，首先需要研究者熟悉蛋白质的多级层次结构。由于氨基酸在三维空间中的不断折叠弯曲，形成了蛋白质的多样性结构蛋白质的结构，一般可分为以下四级结构：

（1）一级结构：组成蛋白质多肽键的氨基酸序列。一级结构是蛋白质最基本的结构，它是由蛋白质序列的氨基酸的一维排列顺序所决定的。蛋白质的一级结构是基础，也决定了蛋白质的二级、三级等结构。

（2）二级结构：指在不同的氨基酸之间的 C=O 和 N-H 基团间的氢键所形成的稳定结构。主要以 α 螺旋和 β 折叠为主。

（3）三级结构：蛋白质二级结构在三维空间中的布局。蛋白质的多肽链在多种二级结构的基础上，在三维空间进一步扭曲旋转所形成具有特定规则的空间结构。

（4）四级结构：亚基与亚基通过相互作用形成的结构，通常，单独的一个肽链会被称为亚基。具有两条或两条以上独立三级结构的多肽链组成的蛋白质，其多肽链间通过次级键相互组合而形成的空间结构称为蛋白质的四级结构。

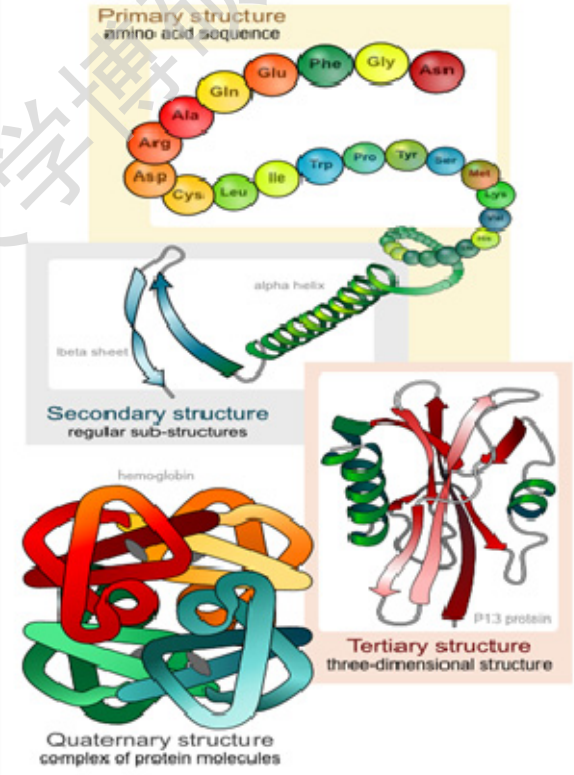

图 1-1 蛋白质一级结构到四级结构的示意图