同济大学
TONGJI UNIVERSITY

博士学位论文

基于机器学习的蛋白质相互作用与功能预测

（国家自然科学基金项目(61173118)资助）

姓　　名：邓磊
学　　号：0810080069
所在院系：电子与信息工程学院
学科门类：工学
学科专业：计算机应用技术
指导教师：关佶红

二〇一二年五月

同濟大學
TONGJI UNIVERSITY

A dissertation submitted to
Tongji University in conformity with the requirements for
the degree of Doctor of Philosophy

Protein Interaction and Function Prediction Based on Machine Learning Techniques

(Supported by the Natural Science Foundation of China (61173118))

Candidate :	Lei Deng
Student Number :	0810080069
School/Department :	School of Electronics and Information Engineering
Discipline :	Engineering
Major :	Computer Application
Supervisor :	Jihong Guan

May, 2012

学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以盈利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年　　月　　日

此为试读，需要完整PDF请访问：www.ertongbook.com

同济大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

摘要

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。

本文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。论文的主要贡献如下：

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，本文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它已有方法具有更高的性能。我们还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，我们还通过多次对负样本(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，我们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它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，我们的方法都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。我们还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，我们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。我们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。实验表明我们的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。

关键词： 蛋白质相互作用，蛋白质相互作用位点，能量热点，蛋白质功能，机器学习，结构比对

Abstract

Proteins are the material basis of all life, the key components of body cells, and play important roles in the process of life activity. The interactions between proteins play important role in the realization of various functions. The study of protein-protein interactions and functions is of crucial importance in understanding the activities of life, disease treatment, drug development and protein design. With the rapid development of molecular biology techniques as represented by high-throughput sequencing, more and more genomes have been sequenced, and also protein sequence and structure data is increasing. Using the traditional experimental methods to identify protein-protein interactions and annotate protein functions can not meet the current demand. Therefore, exploring new computational technologies to predict protein interactions and functions, and uncovering the underlying biological principles have become an increasingly important research topic.

In this thesis, we use machine learning methods to study several important aspects of protein-protein interaction and function prediction: protein interaction site prediction, protein interaction hot spot prediction, protein-protein interaction prediction and protein function prediction. Also, a series of protein-protein interaction and function prediction methods and tools have been developed. The main contents of the paper are as follows:

1. In the area of protein interaction prediction, we have proposed two protein interaction site prediction methods, one is based on ensemble learning and the other is based on structural neighbor templates. We have also developed a semi-supervised method and an ensemble method based on structural neighborhood properties. Furthermore, we have proposed two protein function prediction methods, one is based on sequence composition information and the other is based on structure and multi-data sources.

(1) An effective ensemble-based interaction site prediction method has been developed, which combines bootstrap re-sampling technique, SVM-based fusion classifiers and weighted voting strategy, to overcome the imbalanced problem and effectively uti-

lize a wide variety of features. To improve the usefulness of the proposed method, two special ensemble classifiers are designed to handle the cases of missing homologues and structural information respectively, and the performance is still encouraging. The robustness of the ensemble method is also evaluated by effectively classifying interaction sites from surface residues as well as from all residues in proteins.

(2) Because the protein interface conservation among protein structural neighbors is very significant, we design a protein interface prediction method based on the structural neighbor templates and support vector machine. The known interface residues of the query protein's structure neighbor, have been mapped to the query protein, and the contact frequency map is obtained. Then we use support vector machine to predict interface residues from surface residues. The performance of this method is higher than that of the other existing methods both in DKBM and CAPRI datasets. We have also developed an interactive and usable protein interface prediction web server - PredUS.

(3)Since mutagenesis experiments are expensive and time-consuming, the number of experimental hot spots is very limited. To solve this problem, an iterative semi-supervised algorithm, SemiHS, has been proposed. SemiHS incorporates unlabeled data to overcome the imbalanced problem and to improve the accuracy of the classifier when insufficient training data is available. Also, a new combination of sequence, structure and energy features have been used.

(4)An integrated protein hot-spot prediction method has been developed based on structural neighborhood properties. 108 Euclidean neighbor features and 108 Voronoi neighbor characteristics have been calculated based on the 108 sequence, structure and energy residue features. Then a random forest method is utilized to select the top 46 important characteristics. Due to the imbalance in the protein hot-spot prediction, we also re-sample the negative samples (non-hot spots) to build an integrated classifier, and achieve high prediction performance. On this basis, we also develop protein hot-spot prediction web server - PredHS.

(5)Genome-wide protein-protein interaction prediction algorithm based on the structure. First, for the query protein pairs' structure or homology models, a structure alignment algorithm is used to search structure neighbors. Then we superimpose the complex templates of the structure neighbors to generate an interaction model. A

Bayesian network method is used to evaluate the interaction model. Finally, we use the Naive Bayes method to integrate the structural information and other non-structural information, to build a synthesizing PPI prediction model. Whenever in the benchmark or in the whole genome, our method outperforms the existing non-structure prediction methods and high-throughput experimental methods, and thus has vast application prospect.

2. We have proposed two protein function prediction methods, one is based on sequence composition information and the other is based on structure and multi-data sources.

(1)Four kinds of basic building blocks of protein sequences are investigated, including N-grams, binary profiles, PFAM domains and InterPro domains. The protein sequences are mapped into high-dimensional vectors to predict Gene ontology-based protein function by using the occurrence frequencies of each kind of building blocks. We also demonstrate that the use of feature extraction algorithms such as latent semantic analysis and nonnegative matrix factorization. Experimental results show that protein sequence information can be used to predict protein functions.

(2)Since similar protein structure means similar protein function, we propose a structure-based protein function prediction model - PredGO. Based on the query protein's structure or homolog model, we first use the structure alignment method to search the first-level structure neighbors, then we search the second-level structure neighbors from the structure neighbor database for the query protein's sequence homologys. We design an effective scoring function to evaluate the two-level structure neighbors' annotations, and to annotate the GO terms with higher scores to the query protein. In addition, PredGO use the Bayesian approach to integrate the non-structural information, such as protein sequence and interaction data. Our approach has better prediction accuracy and coverage than the existing non-structural methods, and can be applied to the function annotation of the unknown protein sequences and structures.

Key words: protein-protein interaction, protein interaction sites, hot spots, protein function, machine learning, structure alignment

此为试读,需要完整PDF请访问: www.ertongbook.com

目录

第 1 章 绪论

1.1 研究背景

人类自诞生之日起，就在不断探索生命的奥秘。1838年，德国植物学家施莱登 (Schleiden, M.J.) 和动物学家施旺(Schwann, T.)提出了细胞学说，论证了整个生物界在结构上的统一性，以及在进化上的共同起源。同一年，荷兰化学家Mulder, G.J.对一般的蛋白质进行元素分析发现几乎所有的蛋白质都有相同的实验公式。1865年，遗传学的奠基人奥地利人孟德尔(Mendel, G.H.)在著名的《植物杂交试验》一文中提出了基因的概念，开创了遗传学。二十世纪四十年代，Avery, O.T.等人进行了体外转化实验，证实遗传物质是DNA而不是蛋白质。1953年4月，美国生物学家沃森(Watson, J.D.)和英国物理学家克里克(Crick, F.H.C.)在《Nature》上发表共同研究的成果——DNA分子的双螺旋结构模型[1]。此模型的建立，是分子生物学诞生的标志，打开了“生命之谜”的大门，人类对生命科学的研究进入了基因时代，被称之为“生物学的革命”。在其后的20年中，科学家们逐步地认识了从DNA到蛋白质的编码，掌握了三联密码子的本质。1970年代中期，Sanger发明了末端终止法测序技术[2]，使得全基因组规模的测序成为可能，开启了深入研究生命遗传密码的大门。 1990年10月1日，耗资30亿美元、由全球数百名顶尖科学家组成的研究小组启动了人类基因组研究计划(Human Genome Project)[3]。

1.1.1 生物信息学概述

生物信息学(Bioinformatics)正是随着人类基因组计划的启动而兴起的一门新的交叉学科。它涉及生物学、应用数学和计算机科学，依赖于生物实验和衍生数据的大量储存，依赖于计算机科学和应用数学的基础。从广义上说，生物信息不仅包括基因组信息，如基因的DNA序列、染色体定位，也包括基因产物(蛋白质或RNA)的结构和功能及各生物种间的进化关系等其他信息资源。生物信息学既涉及基因组信息的获取、处理、贮存、传递、分析和解释，又涉及蛋白质组信息学如蛋白质的序列、结构、功能及定位分类、蛋白质连锁图、蛋

白质数据库的建立、相关分析软件的开发和应用等方面，还涉及基因与蛋白质的关系如蛋白质编码基因的识别及算法研究、蛋白质结构、功能预测等，另外，新药研制、生物进化也是生物信息学研究的热点。

对生物分子数据的收集、组织和管理是一切生物信息学研究的基础和出发点。从20世纪80年代开始，涌现了一大批生物信息中心和生物信息数据库。生物信息中心对生物数据库进行维护，并提供服务。国内外重要的生物信息中心包括美国国家生物技术信息中心(NCBI)、欧洲生物信息学研究所(EBI)、欧洲分子生物学实验室(EMBL)、日本国立遗传学研究所(NIG)、中国科学院上海生命科学研究院生物信息中心(BioSino)和北京大学生物信息中心(CBI)。

迄今为止，生物信息数据库总数已经超过500个，而且种类繁多，大体可以分为4个大类，即基因组数据库、核酸和蛋白质一级结构序列数据库、生物大分子(主要是蛋白质)三维空间结构数据库，以及以上述3类数据库和文献资料为基础构建的二次数据库。基因组数据库(如人类基因组数据库GDB、线虫基因数据库ACeDB和酵母基因组数据库SGD等)来自基因组作图，序列数据库(如蛋白质氨基酸序列数据库SWISS-PROT和PIR数据库等)来自序列测定，结构数据库(如蛋白质结构数据库PDB)来自X-衍射和核磁共振结构测定。其中最为著名的为三大综合性核酸数据库(GenBank、EMBL和DDBJ)。三个数据中心各自搜集世界各国有关实验室和测序机构所发布的序列数据，并通过计算机网络每天都将新发现或更新过的数据进行交换，以保证这三个数据库序列信息的完整性。

1.1.2 后基因组时代与蛋白质组学

2003年4月14日，美国人类基因组研究项目首席科学家Francis Collins宣布人类基因组序列图绘制成功, 人类基因组计划的所有目标全部实现。标志着生命科学进入了后基因组时代。在随后的几年中，以454、Solexa和SOLiD为代表第二代高通量测序技术迅速发展，开创了基因组测序的新时代。测序通量呈指数提高, 而成本急剧降低。利用传统Sanger测序法完成的人类基因组计划总计耗资27亿美元，而2008年454生命科学公司(454 Life Sciences)和Baylor医学院基因中心的科学家们利用新一代高通量测序技术为“DNA之父”詹姆斯·沃森(James D. Watson) 测得了首个人类个人全基因组[4]，只花费不到100万美元。2010年，Illumina和ABI先后发布新款测序仪，改进了原有机型，人类进入了数千美元测一个人全基因组的时代。

后基因组时代的生命科学重心，已从解析生命的遗传密码转移到系统研究这些遗传密码代表的生物学功能上来。分子生物学研究发展出了诸如功能 (或结构)基因组学、蛋白质组学、比较基因组学、药物基因组学等重要研究方向。蛋白质组学(Proteomics)一词源于蛋白质(Protein)和基因组(Genome)两个词的结合，是由两位澳大利亚科学家Wilkins和Williams于1994年在第一届国际蛋白质组学专题研讨会上提出。蛋白质组学是在基因组学研究的基础上，大规模、有系统地研究蛋白质的特征及结构，全方位研究细胞内全部蛋白质的动态表达。蛋白质组学可概括为表达蛋白质组学和细胞图谱蛋白质组学，前者利用各种先进技术研究蛋白质表达的整体变化，即研究机体的生长发育、疾病和死亡的不同阶段中，细胞与组织的蛋白质组分的变化；后者主要通过分离蛋白质复合物，系统地研究蛋白质间的相互作用，以建立细胞内信号转达通路的复杂网络图。

与此前的蛋白质研究技术不同，后基因组时代的蛋白质组研究是从整体水平上解析基因组表达的全部蛋白质的结构功能，这就需要有和基因组大规模测序技术类似的高通量研究手段。近年来，随着酵母双杂交(Yeast Two-hybrid)、串联亲和纯化(TAP)、质谱技术(MS)和蛋白质芯片(Protein Chip)等高通量蛋白质组技术的发展，蛋白质序列和相互作用数据迅速增长。截止目前，Swiss-prot数据库中的蛋白质序列数量已经超过53万个。但是，通过生物实验方法确定蛋白质相互作用和功能任然代价较高，数量相对较少。如MIPS，DIP，IntAct，MINT和BioGRID等数据库中实验测定的非冗余相互作用只有约39万个，而具有实验功能标注的蛋白质只有约7万4千个。

1.1.3 蛋白质序列、结构和功能之间的关系

在分子水平上的生命系统演化大部分存在这样的级联规则：基因序列决定氨基酸序列；氨基酸序列决定蛋白质结构；蛋白质结构决定蛋白质功能；选择行为改变等位基因频率(结束循环)。

基因组测序计划产生了越来越多的物种的全基因组DNA序列。为基因组基因序列的识别提供了大量的蛋白质氨基酸数据。在结构基因组计划中，X-射线晶体学(X-ray Crystallography)和核磁共振光谱(NMR Spectroscopy)被用来确定一部分蛋白质的结构，而其它结构则使用同源建模(Homology Modelling)来预测。当代生物信息学的一个重要目标就是对这些序列、结构和功能数据进行收集整理，并研究它们之间的关系。

此为试读,需要完整PDF请访问: www.ertongbook.com