分类号：O221. TP13　　单位代码：10019
密　级：　　学　　号：B06080426

中国农业大学

学位论文

基于核方法的蛋白质-核酸、蛋白质-蛋白质相互作用预测研究

Prediction of Protein-Nucleic Acids and Protein-Protein Interactions Using Kernel Methods

研　　究　　生：邵小健
指　导　教　师：邓乃扬 教授
合 作 指 导 教 师：Gary Bader 教授

申 请 学 位 类 别：管理学 博士
专 业 领 域 名 称：运筹与管理
研　究　方　向：运筹与优化
所　在　学　院：理学院

2011 年 5 月

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国农业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：　　时间：　　年　　月　　日

关于论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。同意中国农业大学可以用不同方式在不同媒体上发表、传播学位论文的全部或部分内容。

(保密的学位论文在解密后应遵守此协议)

研究生签名：　　时间：　　年　　月　　日

导师签名：　　时间：　　年　　月　　日

摘 要

蛋白质在细胞生物过程中起着至关重要的作用，对蛋白质功能的研究是当前生命科学研究的一个重要领域。生物体内蛋白质的功能主要靠它与其他分子的相互作用而实现，其中有蛋白质与DNA/RNA等核酸大分子的相互作用、蛋白质与蛋白质之间的相互作用等。确定DNA/RNA结合蛋白质对研究细胞体内的转录调控、翻译表达等生物过程和构建蛋白质-核酸相互作用网络具有至关重要的作用。在蛋白质结构域层面上确定蛋白质结构域与短肽之间的相互作用则为系统地研究蛋白质与蛋白质之间的相互作用提供更为精细而准确的相互作用信息，有助于更加深刻地理解生物体内的生物过程。用实验方法测定上述蛋白质与核酸、蛋白质与蛋白质之间的相互作用既费时又费力，而且存在无法测定的风险。因此，采用生物信息学的手段预测上述蛋白质与核酸、蛋白质与蛋白质之间的相互作用具有极其重要的理论和实际意义。

目前已有大量的DNA/RNA结合蛋白质数据和蛋白质结构域-短肽相互作用数据，使得采用机器学习的手段构建预测模型变得可行。本文基于现有高通量数据，提出了预测蛋白质与其他生物大分子（核酸分子或蛋白质分子）之间的相互作用的若干计算模型。具体地，主要有以下几个方面的内容：

（1） 对DNA/RNA结合蛋白质的功能注释进行了研究。采用支持向量分类机模型，构建了识别DNA结合蛋白质和RNA结合蛋白质的两种分类机。新模型主要依赖蛋白质序列信息，采用描述蛋白质序列的新方式——“三联体”编码方式对蛋白质进行了特征编码。在蛋白质序列之间的相似度低于25％的非冗余数据集上的不同测试结果表明，DNA结合蛋白分类器和RNA结合蛋白分类器取得了比前人更好或者相当的分类性能。同时，本文还构建了区分DNA结合蛋白与RNA结合蛋白的DNA-RNA结合蛋白分类机，数值试验结果表明了新模型的有效性。本文对描述蛋白质序列的“三联体”特征进行特征选择，选取了可以分别识别DNA结合蛋白和RNA结合蛋白的最为有效的部分“三联体”特征，并发现它们大部分都处于DNA结合蛋白或者RNA结合蛋白的结合表面。这表明“三联体”特征有效地表征了DNA结合蛋白或者RNA结合蛋白在结合表面的结合模式。

（2）对WW结构域-短肽相互作用开展了深入的研究。本文采用“二值正交”编码方式分别对WW结构域和短肽序列进行特征编码，并结合支持向量分类机对WW结构域-短肽对是否发生相互作用进行了预测。本文的试验结果表明仅仅从WW结构域或者短肽配体的序列信息进行预测它们之间的相互作用是行之有效的。此外，本文还通过采用支持向量顺序回归机模型，实现了对WW结构域-短肽相互作用强弱水平的预测，这给生物学家提供了更详尽的信息。本文工作通过构建预测模型初步实现了在整个人类蛋白质组层面上的针对WW结构域和其配体之间相互作用强弱（亲和力）水平的预测。更进一步，我们还设计了预测WW结构域-短肽对是否发生相互作用的网站供研究者使用，具体网址为：http://www.baderlab.org/WWpredictor。

（3）针对定量化预测PDZ结构域-短肽相互作用亲和力开展了研究工作。本文根据现有的首次通过高通量实验获得的关于PDZ结构域-短肽相互作用的具有亲和力的“半量化”数据特性，设计了新的预测模型——半量化支持向量回归机，同时考虑了定量化的正类样本和定性化的负类样本，扩展了传统的仅能考虑定量化数据的支持向量回归机模型。本文提出的新模型不仅能预测PDZ结构域-短肽序列之间是否发生相互作用，也能对发生相互作用的PDZ结构域-短肽对预测其

亲和力。本文的结果表明了仅从PDZ结构域和短肽序列的氨基酸序列信息对它们之间相互作用的强弱进行预测是可行的。同时也表明整合了负类样本信息的模型能有效地提高传统回归模型的预测性能。此外，试验结果还表明了本文提出的新模型能准确地预测对短肽序列进行单点变异之后的新短肽序列与PDZ结构域之间相互作用的亲和力的大小。本文的工作为从事PDZ结构域-短肽相互作用亲和力的研究提供了非常有用的工具和新的视角，也为研究其他结构域-短肽相互作用亲和力的工作提供了借鉴。具体的Matlab程序可从http://baderlab.org/Data/PDZAffinity下载。

本文还就与本论文研究密切相关的一些未来将要从事或者可能从事的研究工作进行了展望和讨论。

关键词：蛋白质功能 结构域–短肽相互作用 定量化预测 支持向量机

Abstract

Proteins play key roles in almost all cellular processes. Studies on protein functions are currently an important field of life science research. Proteins carry out most of their functions *in vivo* mainly by their interactions with other molecules, such as protein-nucleic acid interactions, protein-protein interactions, etc. DNA/RNA-binding proteins play key roles in transcriptional regulation, transcriptional expression and many other biological processes. Determining these proteins also helps to construct the protein-nucleic acid interaction network. Determining protein domain-peptide interactions at domain level can provide more precise and accurate information to systematically study protein-protein interactions and thus help to understand more detailed biological processes *in vivo*. Experimental determination of these protein-nucleic and protein-protein interactions is not only labor intensive but also time consuming, and even it is unable to be detected. Accordingly, it would be highly desirable to develop *in silico* approaches to predict protein-nucleic acid interactions and protein-protein interactions with both of theoretical and practical significance.

With a fairly large set of DNA/RNA-binding proteins and high-throughput experimental technique produced protein domain-peptide interaction data becoming available, it is possible for us to develop machine-learning-based computational methods to predict them. In this thesis, we propose computational models to predict DNA/RNA-binding proteins and protein domain-peptide interactions. The detailed results are summarized as follows:

(1). We design DNA/RNA-binding protein classifiers to recognize DNA/RNA-binding proteins from non-nucleic-acid-binding proteins via support vector machines. All of the proteins are encoded by conjoint triad features which extract information directly from primary sequence of proteins. Both self-consistency and jackknife tests show promising results on the protein datasets in which the sequences identity is less than 25%. We also construct a classifier to differentiate DNA-binding proteins from RNA-binding proteins, and the validation results show the efficiency of our new model. The improvement suggests our new models might provide a complementary tool to existing sequence-based prediction methods. In addition, we use mRMR feature selection method to select the most informative conjoint triad features for both DNA-binding and RNA-binding protein classifiers. Further investigations find that these informative conjoint triad features lie in the binding surface of the proteins, and this may suggest that conjoint triad features appropriately capture the binding patterns in the binding surface of the proteins.

(2). We first construct a binary prediction model to predict WW domain-peptide interactions. We encode both WW domain sequences and peptides by “orthogonal binary” method, and apply support vector machines to predict whether they interact or not given both WW domains and peptides. Our experimental results show it is possible to predict binary interactions between WW domains and peptides from their sequences, and predicted potential peptides can be treated as important resources that will facilitate biologists for their further experiments. We then apply support vector ordinal regression model to predict which binding level a WW domain-peptide interaction pair belongs to.

Different from only predicting whether they interact or not, the new model can tell which binding levels they belong to. It is more interesting since the predicted binding levels could provide more detailed information to biologists. In summary, this work explores computational prediction models trained on large-scale experimental data to predict WW domain-peptide interactions as well as the binding strength levels in human proteome. The web tool for predicting human WW domain-peptide interactions can be accessed freely at http://www.baderlab.org/WWpredictor.

(3). We establish a new regression framework to quantitatively predict PDZ domain-peptide interactions. It is important to consider the binding strength of domain-peptide interactions to help us to construct more biologically relevant protein interaction networks that consider cellular context and competition between potential binders. Based on a "semi-quantitative" data which considers both positive (quantitative) and negative (qualitative) interaction data for mouse PDZ domains, we develop a novel regression model named "SemiSVR" to quantitatively predict interactions between PDZ domains and their peptide ligands using primary sequence information. The proposed new model can not only predict which PDZ domain-peptide pairs are likely to interact but also can predict affinity of PDZ domain-peptide interactions. We show that it is possible to learn from existing quantitative and negative interaction data to infer the relative binding strength of interactions involving previously unseen PDZ domains and/or peptides given their primary sequence. We find that incorporating negative data improves quantitative interaction prediction. In addition, we show that our method can correctly predict the direction and relative magnitude of affinity changes in the mutant ligand compared with the wild type. In a word, our work provides a very useful tool for quantitatively predicting PDZ domain-peptide interactions and new insights, and also will serve as a reference for studying quantitative prediction on other domains. The Matlab codes for our SemiSVR predictor are available freely at http://baderlab.org/Data/PDZAffinity.

Key words: Protein functions, Protein domain-peptide interactions, Quantitative predictions, Support Vector Machines

目 录

表格

图

第一章 引言

生物信息学是一门新兴的交叉学科，其涵盖了生物学、信息科学、统计学、数学等诸多学科，它不仅是当今生命科学和自然科学的重大前沿领域之一，同时也将是21世纪自然科学的核心领域之一。从国外近几年的研究和应用情况来看，生物信息学在理论上促进了生物学（特别是分子生物学）的发展，使人类对生命本质的认识更加深刻。随着后基因时代的到来，越来越多的生物学数据呈现在生物学家的面前，如何利用这些数据得到有趣的生物学知识是生物信息学家一直研究的重点。其中随着人类基因组计划的完成，得到的大多数基因的产物为相应的蛋白质，要认识基因的功能，就需要研究其对应的蛋白质。目前，对蛋白质功能的预测是生物信息学研究中的一个重要课题，而蛋白质发挥其功能主要是靠其与其他分子或者蛋白质相互作用而实现的。近年来，随着高通量生物数据的获得，人们已经不仅仅只停留在从单个蛋白或个体的孤立情况下看生物学问题，如何从系统的角度看待生物学问题变得尤为重要，系统生物学一个新的研究体系应运而生。

在系统生物学中，各种生物学相互作用网络是研究的重点。其中主要有关于蛋白质与DNA/RNA等核酸大分子的相互作用、蛋白质与蛋白质之间的相互作用网络等。DNA结合蛋白质调节一些非常重要的细胞生物过程，比如转录调控、复制、DNA修复、重组及其他在细胞发育过程中较为重要的环节。RNA结合蛋白质则在基因表达不同阶段与RNA相互作用共同调控蛋白质的合成。蛋白质不仅与DNA/RNA等核酸大分子进行相互作用，同时也与其他蛋白质进行相互作用，在特定的时间和空间内完成特定的功能。研究蛋白质与蛋白质之间的相互作用将有助于了解细胞中不同生命活动之间的相互关系。而蛋白质与蛋白质之间的相互作用又主要通过蛋白质中某些特定功能的结构域与结构域相互作用和结构域与短肽相互作用来实现。在结构域层面上研究蛋白质与蛋白质之间的相互作用将为系统研究蛋白质相互作用网络提供更为精细而准确的相互作用信息，有助于更加深刻地理解生物体内生物过程。

1.1 生物学背景知识简介

1.1.1 核酸

核酸是由许多核苷酸聚合而成的生物大分子化合物，为生命的最基本物质之一[1, 2]。根据化学组成以及核苷酸的排列顺序不同，核酸可分为脱氧核糖核酸（Deoxyribonucleic acid），简称DNA和核糖核酸（Ribonucleic acid），简称RNA。DNA和RNA都是由一个个的核苷酸排列组合而成。而核苷酸主要由碱基、戊糖和磷酸三部分组成。DNA和RNA在核苷酸的组成上略有不同。

DNA是生物体内传递遗传信息的主要物质。DNA大分子中含有的戊糖为脱氧核糖。DNA对应含有四种碱基，分别为腺嘌呤（adenine，A）、鸟嘌呤（guanine, G）和胞嘧啶（cytosine, C），胸腺嘧啶（thymine，T）。DNA大部分为双链分子，两条单链反向平行地按照碱基互补配对原则（A与T、C与G配对）结合在一起，呈双螺旋状。

RNA在蛋白质合成过程中起着重要作用。与DNA不同，RNA中含有的戊糖对应为核糖。虽然RNA也对应含有四种碱基，但是其组成不尽相同。与DNA一样，RNA中也含有腺嘌呤（adenine，A）、鸟嘌呤（guanine, G）和胞嘧啶（cytosine, C），但是没有胸腺嘧啶（thymine, T），却含有尿

嘧啶（uracil, U）。绝大部分RNA分子都是线性单链状结构，但是有时也会通过碱基互补配对（A与U、G与C配对）形成局部的二级结构和三级结构。

在生物体内发现的RNA的种类较多，主要有信使核糖核酸（messager RNA），简称mRNA；转移核糖核酸（transfer RNA），简称tRNA；和核糖体的核糖核酸（ribosome RNA），简称rRNA等。除上述3种主要的RNA外，还有一种小核RNA（small nuclear RNA），简称snRNA。

近年来，由于新技术的不断发展，人们也发现了许多非编码RNA（non-coding RNA，ncRNA）。例如，其中微RNA（microRNA，miRNA）是一种具有发夹结构的非编码RNA，长度一般为21-25个核苷酸，存在于所有真核生物中，在真核基因表达调控中起着重要的作用，它可以与靶mRNA结合，产生转录后基因沉默作用（post-transcriptional gene silencing, PTGS）[3]。

目前，核酸研究的进展日新月异。人们对DNA和各种各样的RNA的研究不断深入，揭示了更多生命的奥秘，使人们对生命本质的认识更加深入。同时，为了便于人们进行进一步的研究，人们建立了专门的数据库来存储核酸数据。其中，常用的核酸数据库主要有：

美国的核酸数据库GenBank，网址为http://www.ncbi.nlm.nih.gov；欧洲核酸序列数据库EMBL，网址为http://www.embl-heidelberg.de；日本核酸序列数据库DDBJ，网址为http://www.ddbj.nig.ac.jp。

1.1.2 蛋白质

蛋白质（Protein）是由不同氨基酸组成的具有一定空间结构的生物大分子物质，在几乎所有的生物过程中起着关键作用[1, 4]。组成蛋白质的基本单位是氨基酸。目前主要发现有20种“标准”氨基酸，这些氨基酸由不同的三个核苷酸组合（也称“遗传密码子”）所编码构成，且部分氨基酸可由多个不同的密码子编码。为方便后续的分析，表1-1列出了20种氨基酸的名字，更为具体的20种“标准”氨基酸对应的密码子编码和相应的化学分子式等信息可见[4]。

表1-1 20种氨基酸汇总

Amino Acid	氨基酸	3-Letter	1-Letter
Alanine	丙氨酸	Ala	A
Arginine	精氨酸	Arg	R
Asparagine	天冬酰胺	Asn	N
Aspartic acid	天冬氨酸	Asp	D
Cysteine	半胱氨酸	Cys	C
Glutamic acid	谷氨酸	Glu	E
Glutamine	谷氨酰胺	Gln	Q
Glycine	甘氨酸	Gly	G
Histidine	组氨酸	His	H
Isoleucine	异亮氨酸	Ile	I
Leucine	亮氨酸	Leu	L
Lysine	赖氨酸	Lys	K
Methionine	甲硫氨酸	Met	M
Phenylalanine	苯丙氨酸	Phe	F
Proline	脯氨酸	Pro	P

Serine	丝氨酸	Ser	S
Threonine	苏氨酸	Thr	T
Tryptophan	酪氨酸	Trp	W
Tyrosine	色氨酸	Tyr	Y
Valine	缬氨酸	Val	V

注：表中各列依次给出了20种氨基酸的英文名字、中文名字及其相应的3字母缩写和单字母缩写。

蛋白质的氨基酸序列由对应基因的核苷酸序列所决定。蛋白质的氨基酸序列经过折叠可以形成不同的结构。通常，蛋白质的分子结构可以有以下几种形式。一级结构：蛋白质多肽链中氨基酸的排列顺序，也即蛋白质序列。二级结构：蛋白质分子局部区域内，多肽链沿一定方向盘绕和折叠的方式，有α螺旋、β纸片和coil或者loop区等。三级结构：蛋白质的二级结构基础上借助各种次级键卷曲折叠成特定的球状分子结构的空间构象。四级结构：多亚基蛋白质分子中各个具有三级结构的多肽链，以适当的方式聚合所形成的蛋白质的三维结构。

目前常用的蛋白质数据库主要有：1) 蛋白序列数据库：PIR数据库、Uniprot（http://www.uniprot.org）和 SWISS-PROT（http://www.ebi.ac.uk/swissprot/）；2) 蛋白质三维结构数据库：PDB（http://www.rcsb.org/pdb/）。

1.1.3 结构域

结构域是在蛋白质三级结构内的独立折叠单元，通常都是几个超二级结构单元的组合[1, 5]。

结构域（Structural Domain）是蛋白质结构、功能和进化的基本单位，是由蛋白质序列中的一个局部片段紧密折叠而成的，独立于序列中的其他氨基酸残基，并具有较为稳定的三级结构。较大的蛋白质分子往往由两个或多个相对独立的结构域构成。对于较小的蛋白质分子来说，结构域和它的三级结构往往是一个意思，也就是说这些蛋白质是单结构域。同一个蛋白质结构域可以出现在多个蛋白质分子中。蛋白质结构域作为功能单元往往以不同的方式组合出现在不同的蛋白质分子中以行使不同的功能。结构域自身是紧密装配的，但结构域与结构域之间关系松懈。结构域与结构域之间常常由一段长短不等的肽链相连，形成所谓loop区。不同蛋白质分子中结构域的数目不同，同一蛋白质分子中的几个结构域彼此相似或很不相同。常见结构域的氨基酸残基数在25～500个之间，最小的结构域只有40～50个氨基酸残基，大的结构域可超过400个氨基酸残基。

目前，常用的注释蛋白质结构域的数据库主要有：

Pfam[6]：Pfam数据库是一个蛋白质家族的大集合，这些蛋白质家族主要是根据其包含的结构域进行分类，依赖于蛋白质多序列对比和隐马尔科夫模型（HMMs）。具体网址为http://pfam.janelia.org/，在不同的国家设有镜像。目前的版本：Pfam 25.0（2011年3月, 共含12273个蛋白家族）。Pfam数据库有两个组成部分：Pfam-A和Pfam-B。Pfam-A数据库的数据质量比较高，都是经过人工筛选的。Pfam-B数据库是基于计算预测方法得到的蛋白质结构域数据库。该数据库虽然质量较低，但其可以被用来鉴别功能保守区域，尤其是对于那些还没有被Pfam-A数据所收录的蛋白质数据。

Smart[7]：简单模块结构搜索工具（Simple Modular Architecture Research Tool，SMART），可以进行蛋白质结构和功能分析，具体网址为http://smart.embl-heidelberg.de/。可以说是蛋白质结构

预测和功能分析的工具集合。简单点说，就是集合了一些工具，可以预测蛋白质的一些二级结构，如跨膜区（Transmembrane segments）、复合螺旋区（coiled coil regions）、信号肽（Signal peptides）和蛋白质结构域（PFAM domains）等。

CDD[8]：NCBI的蛋白质保守结构域数据库（Conserved Domain Database，CDD），是一个蛋白质注释数据库，该数据库收录了大量保守结构域序列信息和蛋白质序列信息。检索者通过CD-Search服务，可获得蛋白质序列中所含的保守结构域信息，从而分析、预测该蛋白质的功能。具体网址为：http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi。

ProDom[9]：ProDom是建立在SWISS-PROT数据库基础上的蛋白质结构域数据库（protein domain database，ProDom）。其原理是基于递归的PSI-BLAST检索，由SWISS-PROT蛋白质序列库中探查到的同源结构域组成。网址为http://prodom.prabi.fr/prodom/current/html/home.php。

1.2 蛋白质与其他大分子相互作用及其研究意义

相互作用组学属于系统生物学的范畴，是系统生物学中的一个重要研究内容。相互作用组学通常定义为细胞体内各种分子相互作用的总称，主要研究蛋白质与核酸相互作用网络（如转录因子-基因调控网络）和蛋白质与蛋白质相互作用网络等[10]。本节将主要介绍蛋白质与核酸相互作用网络和蛋白质与蛋白质相互作用网络的内容。

1.2.1 蛋白质与核酸分子相互作用

如前所述，核酸和蛋白质是生物体内主要的大分子，各自具有其特定的分子结构和生物学功能。核酸是基本的遗传物质，而蛋白质是生命的物质基础，贯穿生命的整个生理过程。但是它们各自行使功能的过程在一定程度上受到它们之间的相互作用影响。也就是说，细胞中各种重要的生理过程，诸如细胞的生长、繁殖、遗传和代谢，细胞内部的信号传导等，都是以蛋白质与核酸之间的相互作用为纽带的。

目前涉及蛋白质-核酸相互作用的重要数据库主要为蛋白质-核酸识别数据库，具体网址为http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/recognition.html。该数据库涵盖了蛋白质-核酸复合物数据库，核苷酸-蛋白质相互作用数据库，以及蛋白质-核酸相互作用的热力学数据库（ProNIT）[11]等。这些数据为研究蛋白质与核酸之间的相互作用关系提供了丰富的信息。

1.2.2 蛋白质与蛋白质分子相互作用

随着2000年酵母大规模蛋白质相互作用网络图谱的成功描绘，蛋白质相互作用特别是大规模蛋白质相互作用研究成为生命科学领域的又一个研究热点[12-15]。蛋白质在生物体内行使其功能往往并不是单一完成的，而是通过与其他蛋白质的相互作用来完成。了解蛋白质与蛋白质分子相互作用对人们理解蛋白质如何行使其功能和理解整个细胞过程有着非常重要的作用。蛋白质与蛋白质之间的相互作用类型繁多，各有特点。同时这些复杂的相互作用之间却又遵循着某些特定的规律。可以想像，大多数情况下，蛋白质与蛋白质相互作用往往是通过一个或者若干个局部功能区域的结合而实现的。而蛋白质结构域就是蛋白质行使生物学功能的主要功能模块。过去十几年

的研究发现，蛋白质与蛋白质相互作用主要分为两大类：蛋白质结构域与结构域相互作用，和蛋白结构域与其识别的另一个蛋白的短肽序列相互作用。此外，也包含了蛋白结构域与含修饰后氨基酸位点（常见的为磷酸化、乙酰化位点等）的短肽相互作用等，具体可参见图1-1。结构域-结构域之间的相互作用主要由来自两个不同蛋白质各自的全局结构域接触形成，该相互作用比较稳定，且具有较大的相互作用表面（基于现有的三维结构信息，结构域-结构域之间相互作用的表面平均大致在2.0Å^{2}左右）。而结构域-短肽相互作用则主要是由含有全局结构域的蛋白质去识别另一个含有线性短肽序列的蛋白质，该相互作用往往是瞬时的，具有较小的相互作用表面。通常这种类型的相互作用主要发生在信号传导和调控网络中。而正因为该种相互作用通常是瞬时的，要对它们进行实验测定显得更为困难[16]。

为了便于人们更好地利用现有实验获得的数据进行系统的分析和比较，有必要对各种不同实验获得的数据加以处理和整合，并进行适当的存储。下面简单地介绍一下目前使用较多、数据信息较为完善的公共数据库。大部分的蛋白质-蛋白质相互作用数据库存放通过实验获得的数据，也有小部分存放计算预测得到的数据，或者两种类型的皆有。具体的主要有：

DIP（Database of Interacting Proteins）[17]：该数据库收集了由实验确定的蛋白质-蛋白质相互作用对。网址为：http://dip.doe-mbi.ucla.edu/dip/。

BIND[18]：该数据库收集了至少在一篇文献里提到的实验验证的蛋白质-蛋白质相互作用对。该数据库是较早的专业收录文献记载的蛋白质相互作用数据库，目前是BOND分子相互作用数据库的一个子数据库。网址为：http://bond.unleashedinformatics.com/Action。

MINT（Molecular Interactions Database)[19]：该数据集收集了由专家收集的实验验证的蛋白质-蛋白质相互作用对。该数据库网址为：http://mint.bio.uniroma2.it/mint/Welcome.do。

MIPS（Munich Information Center for Protein Sequences）：德国慕尼黑蛋白质数据信息中心数据库主要收集基于不同基因组的蛋白质序列数据的数据库。该数据库包括了MPact，提供了关于*S.cerevisiae*的相互作用数据[20]；而MPPI是唯一一个经过专家手工挑选的高质量的关于哺乳动物的蛋白质-蛋白质物理相互作用数据库[21]；此外还包括了Negatome数据库专门收集不太可能直接发生物理相互作用的蛋白质对的数据[22]。网址为：http://www.helmholtz-muenchen.de/en/ibis。

BioGrid[23]:该数据库整合了不同物种间的蛋白质-蛋白质分子之间的物理相互作用和遗传相互作用（Genetic Interaction）数据库，该数据库也收集了经实验验证的蛋白质相互作用对。具体网址为：http://thebiogrid.org/。

此外，也有部分基于计算模型得到的“假定”（putative）蛋白质-蛋白质相互作用数据库，如PRISM、OPHID和3D-partner等[24-26]。

上述蛋白质数据库给研究者们提供了各种各样不同的选择。研究者可以根据不同的需要选择特定的蛋白质数据库，也可以整合上述多个蛋白质相互作用数据库进行分析。

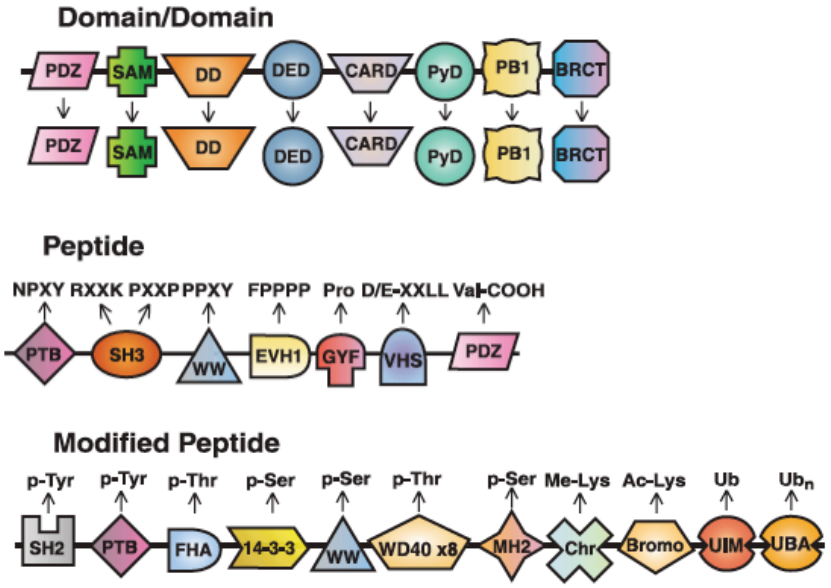


图1-1 蛋白质结构域相互作用模式

上图为蛋白质结构域-结构域相互作用图示。中图为结构域-短肽相互作用图示。下图代表了部分结构域识别含翻译后修饰的氨基酸的短肽序列。

下面分别介绍一下结构域与结构域相互作用和结构域与短肽相互作用。

1. 结构域与结构域相互作用

如前所述，绝大多数的蛋白质的结构和功能单位是结构域，它们调节着蛋白质之间的相互作用。深入研究蛋白质结构域与结构域之间的相互作用可以更加深刻地理解蛋白质与蛋白质分子之间的相互作用机制，也可以帮助用计算的方法来推断全基因组上的蛋白质与蛋白质之间的相互作用，进而构建蛋白质相互作用网络。

目前常见的蛋白质结构域与结构域相互作用数据库主要有：

iPfam和3did是两个主要从PDB数据库中获取的关于结构域-结构域之间相互作用的数据库[27, 28]。这两个数据库各自的网址分别为：http://www.sanger.ac.uk/Software/Pfam/iPfam/ 和 http://3did.irbbarcelona.org。

DOMINE[29]整合了上述两个基于蛋白质三维结构的数据库，同时也包括了许多由若干个主流的生物计算模型推断得到的蛋白质结构域-结构域相互作用数据。该数据库[30]目前包含了大约26219个不重复的结构域-结构域相互作用对，涵盖了5140个Pfam定义的结构域，是研究蛋白质结构域相互作用领域的一个有效的数据资源。具体网址为：http://domine.utdallas.edu。

与DOMINE比较类似的一个数据库为DIMA[31]，该数据库也同时包括了基于蛋白质三维结构获得的结构域相互作用对和由计算模型预测得到的结构域相互作用对。不同之处是该数据库还通过Negatome数据库筛选过滤了部分假阳性相互作用对。此外该数据库还提供了针对各个不同结构域的相互作用网络进行可视化的功能。具体网站为：http://webclu.bio.wzw.tum.de/dima。

上述数据库之间存在着或多或少的数据重合，但也各具优势，为研究人员提供了重要的数

据资源，并为他们研究所需要的数据采集提供了便利。此外，InterDom数据库[32]也涵盖了一些结构域-结构域相互作用对，但是该数据库2007年后尚未有进一步的更新，所以已经逐渐被其他数据库所替代。

2. 结构域与短肽相互作用

蛋白质结构域除了与其他结构域发生相互作用外，很多蛋白质结构域也与其他蛋白质短肽配体发生相互作用。这部分结构域通常涉及到信号传导和调控网络过程。本节图1-2给出了部分常见的短肽识别的结构域（或者称模块识别结构域）。图1-3给出了蛋白质结构域-短肽相互作用对的一个图示。图中左侧蛋白质A和B发生相互作用，可能是因为SH3或者WW结构域和含脯氨酸的短肽序列相互作用，也可能是因为PDZ结构域与另一蛋白质C末端序列发生相互作用。更加确切地知道蛋白质的哪部分区域发生相互作用可以帮助人们更深入地理解生物内在的蛋白质相互作用机制，同时也有利于人们对某些疾病病理的研究。目前研究最多的蛋白质结构域-短肽相互作用对主要有涉及信号传导的SH2、SH3结构域-短肽相互作用、PDZ结构域-短肽相互作用和WW结构域-短肽相互作用等。

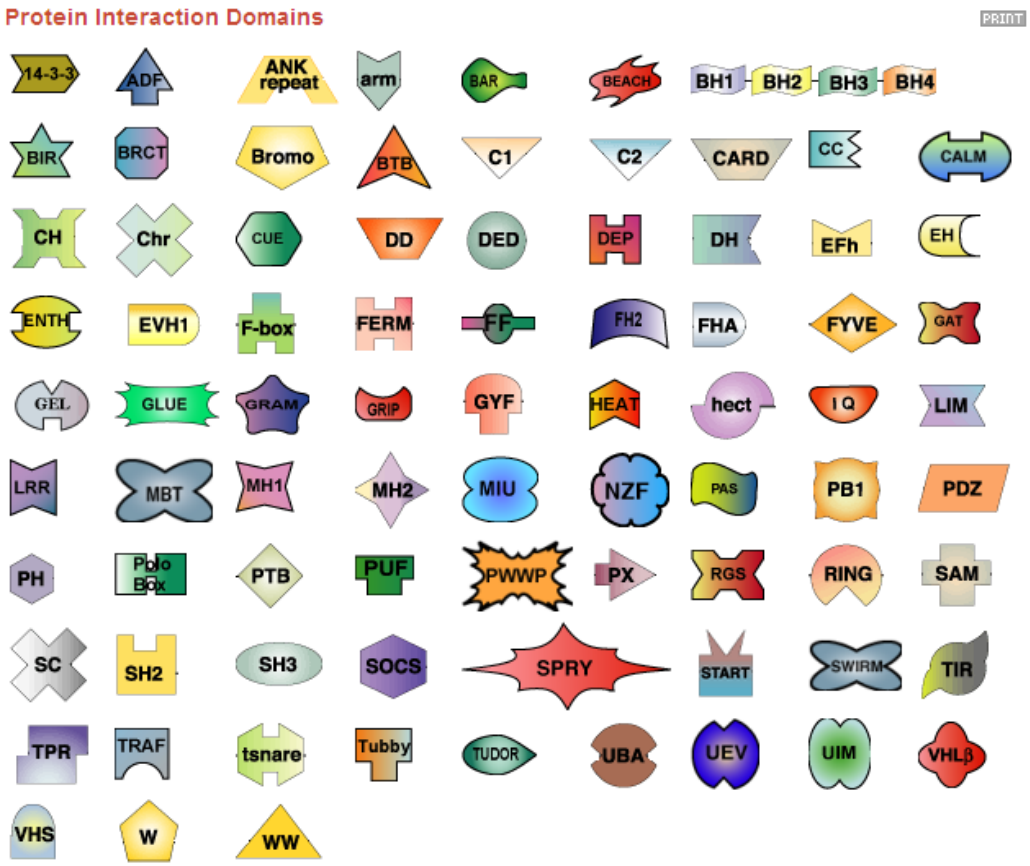


图1-2 模块识别结构域图示（摘自http://pawsonlab.mshri.on.ca/index.php）

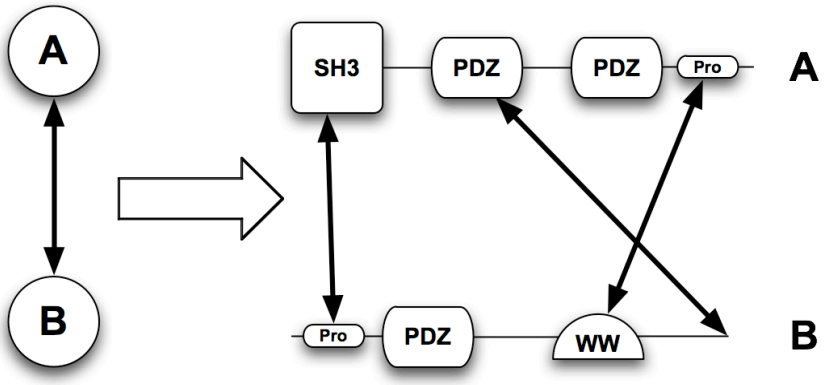


图1-3 蛋白质-蛋白质相互作用由结构域-短肽相互作用完成

蛋白质A和B发生相互作用，可能是因为SH3或者WW结构域和含脯氨酸的短肽序列相互作用，也可能是因为PDZ结构域与另一蛋白质C末端序列发生相互作用。

类似与构建蛋白质结构域-结构域相互作用数据库，研究人员也构建了许多专门记录结构域-短肽相互作用的数据库，具体的有：

ELM[33]：该数据库收集了大量文献中报导过的短肽模体序列及其相互作用结构域数据。该数据库的网址是：http://elm.eu.org/。

DOMINO[34]：该数据库存放了大约超过3900对由实验获得的结构域与短肽相互作用对。其中涉及的结构域包含了SH3, SH2, 14-3-3, PDZ, PTB, WW, EVH, VHS, FHA, EH, FF, BRCT, Bromo, Chromo和GYF等常见的结构域。该数据库的网址为http://mint.bio.uniroma2.it/domino/。

ADAN[35]：该数据库收集了大约超过3505对基于蛋白质三维结构的结构域-短肽相互作用对。其中涉及的三维结构数据中大约有42.6%来自于高质量（high-resolution）的X射线实验得到的数据，15.8%的数据来自与NMR得到的数据，和大约41.6%的由预测模型得到的三维结构的数据。该数据库主要包含了14-3-3, BRCT, FHA, PDZ, PH, Polo box, PTB, PTPc, RA, RBD, SH2, SH3, UBQ, VHS, WD40, WW, ARM, FF, MH2, TRP 等结构域。该数据库的网址为http://adan-embl.ibmc.umh.es/。

PepX[36]: 该数据库从PDB数据库的1431个蛋白质复合体中获取了505对非冗余的蛋白质-短肽相互作用蛋白质复合体。该非冗余数据集是通过对1431个蛋白质复合体根据复合体中结合表面（interface）的模式进行聚类获取的，是目前数据最全的、涵盖最多种结构域-短肽相互作用结构模式的非冗余数据集。该数据库的网址为http://pepx.switchlab.org。

此外，Phospho.ELM (http://phospho.elm.eu.org)[37]是一个专门收集含有磷酸化丝氨酸、苏氨酸和酪氨酸的短肽序列与SH2结构域相互作用的数据库。PDZbase[38]是一个专门存放由实验确定的关于PDZ结构域与短肽配体之间相互作用的数据库。

1.3 若干支持向量机模型简介

本文采用支持向量机方法作为研究蛋白与其他大分子相互作用预测的模型。针对各个具体的生物问题，将会涉及到不同的模型。此节我们将简单介绍本文涉及的各种支持向量机模型及其变

形。

支持向量机（Support vector machine, SVM）最初是由Vapnik等人研究小样本统计学习理论得到的针对分类问题的一个机器学习模型[39-42]。因其在处理小样本数据具有较强的泛化性能，同时有效的避免了“维数灾难”问题而引起了越来越多学者的关注。随后，支持向量机被逐步推广到回归问题[43]，多分类问题[44]和半监督学习问题[45]等。

1.3.1 支持向量分类机简介

设给定训练集为 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，其中 $x_i=(x_{i1},...,x_{in})^T\in R^n$ 为输入，$y_i\in\{+1,-1\}$ 为输出，是x_i对应的类别标号，$i=1,2,...,l$。标准线性可分支持向量机的目标是找到一个可以区分这两类训练点的具有最大间隔的分类面。以如图1-4所示的二维空间(n=2)上的线性可分分类问题为例。有许多直线能将这两类点（“■”和“o”两类）正确分开，如图中蓝色虚线。然而，最好的直线应该是具有最大“间隔”的那条直线（图中绿色实线）。假定该直线为：$(w\cdot x)+b=0$。最大化对应“间隔”的思想，就导致了如下的最优化问题：

$$\min_{w,b}\quad \frac{1}{2}\|w\|^2, \tag{1-1}$$

$$s.t.\quad y_i((w\cdot x_i)+b)\geq 1,\quad i=1,\cdots,l \tag{1-2}$$

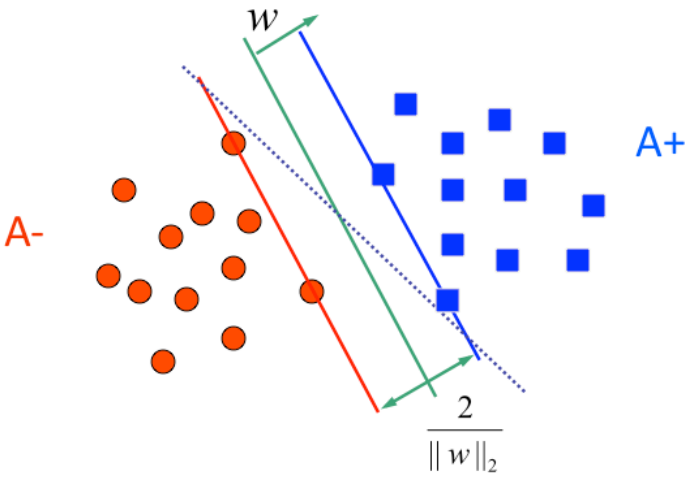


图1-4 线性可分问题

（基于最大间隔的支持向量分类机模型）

对于线性不可分问题，支持向量分类机通过引入“松弛变量”$\xi\geq 0$加以解决，对应如下最优化问题：

$$\min_{w,b,\xi}\quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i, \tag{1-3}$$

$$s.t.\quad y_i((w\cdot x_i)+b)\geq 1-\xi_i,\quad i=1,\cdots,l \tag{1-4}$$

$$\xi_i\geq 0,\quad i=1,\cdots,l \tag{1-5}$$

其中C为惩罚参数，权衡模型的经验风险与推广能力。

针对非线性分类问题，支持向量分类机运用“核函数”技巧，将原始空间的样本点映射到“高维”空间（或者Hilbert空间），并在该高维空间中应用线性分划模型来获取分类超平面。常用的核函数有线性核函数、Gauss径向基核函数、多项式核函数等。

通常，支持向量分类机（C−SVC）不直接求解其原始最优化问题，而是通过求解它的对偶问题来求得原始问题的解。非线性分类问题之对应的对偶问题如下：

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_jK(x_i\cdot x_j)\alpha_i\alpha_j-\sum_{j=1}^{l}\alpha_j, \tag{1-6}$$

$$s.t.\quad \sum_{i=1}^{l}y_i\alpha_i=0, \tag{1-7}$$

$$0\leq\alpha_i\leq C,\quad i=1,\cdots,l \tag{1-8}$$

求解上述对偶问题得到最优解$\alpha^*=(\alpha_1^*,\cdots,\alpha_l^*)^T$，并由此构造决策函数：

$f(x)=\mathrm{sgn}(g(x))=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^*y_iK(x_i,x)+b)$。

1.3.2 支持向量回归机简介

支持向量回归机是针对回归问题而提出的[43]。与分类问题中的样本类别为离散的数值如｛+1，-1｝不同，回归问题中的样本对应的输出Y值是取连续的值。通过引入“ε−不敏感损失函数”，可以构建支持向量回归机模型。本文就标准的支持向量回归机进行简要介绍。

给定训练集$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，其中$x_i=(x_{i1},...,x_{in})^T\in R^n$为输入，$y_i\in R$为输出，是$x_i$对应的回归值，$i=1,2,\cdots,l$。线性支持向量回归机寻求具有在“$\varepsilon$带”允许范围内的训练误差最小并且最为平坦的回归函数：$f(x)=(w\cdot x)+b$。这就得到了如下的回归模型：

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}L_\varepsilon(y_i,f(x_i)) \tag{1-9}$$

其中C为参数，调节回归模型的正则项与经验误差之间的权衡。而$L_\varepsilon(y,f(x))$即为“ε−不敏感损失函数”，并且具有如下形式：

$$c(x,y,f(x))=\begin{cases}0, & |y-f(x)|<\varepsilon;\\ |y-f(x)|-\varepsilon, & otherwise\end{cases} \tag{1-10}$$

“ε-不敏感损失函数“的含义是，当x点的观察值y与预测值$f(x)$之间的误差不超过事先给定的ε时，

则认为该点的预测函数值*f(x)*是无损失的，尽管预测值*f(x)*和观察值*y*可能并不完全相等。

回归问题也有线性和非线性之分。针对非线性回归问题，支持向量回归机模型也采用“核函数”技巧将原始训练点映射到高维空间，然后再在高维空间中应用线性支持向量回归机模型。

1.3.3 支持向量顺序回归机简介

支持向量顺序回归机是标准支持向量机的一个推广，它是针对一类特殊的多分类问题而提出的。通常的多分类问题（假设记为*k*类）的类与类之间是没有顺序的，而此处处理的多分类问题的类与类之间从1到*k*是有顺序的，换言之，第*j*类和第*j*-1，*j*+1类相邻，而第*j*-1和第*j*+1类之间并不相邻。通过直接应用最大间隔的思想，Shashua等人提出了基于固定间隔和总和间隔两种不同形式的支持向量顺序回归机模型[46]，之后Chu等人[47, 48]针对文[46]中并没有考虑各个类别对应的阈值*b*之间的大小关系，对其做出了改进。本文就Chu等人提出的最新模型进行简要的介绍，并以3类分类问题为例加以说明。

设训练集$T=\{x_1^1,\cdots,x_{l^1}^1,x_1^2,\cdots,x_{l^2}^2,x_1^3,\cdots,x_{l^3}^3\}$，其中$x_i^j\in R^n$为输入，表示第*j*类中的第*i*个样本点，$j=1,2,3$为相应样本点的类别标号，$l^j$为第*j*类样本点的个数。3类支持向量顺序回归机的目标是找到2个相互平行的分类超平面$(w\cdot x)=b_j, j=1,2$，并且使得两个相邻的类之间的间隔最大，其中$w\in R^n$，$b_1 \leqslant b_2$。它们将整个空间分成3个区域，支持向量回归机将根据样本点所在的区域来推断其类别归属，也即对应的决策函数为

$$f(x)=\min_{j\in\{1,2,3\}}\{j:(w\cdot x)-b_j<0\}, \tag{1-11}$$

其中$b_3=+\infty$。

对于每一个训练样本点x_i^k，考虑其相对各个分划面的“经验误差”ξ_{ki}^j（针对上界误差，$k\le j$）和ξ_{ki}^{*j}(针对下界误差，$k>j$)，如图1-5所示。为了便于统一写出最终模型，引入辅助变量$b_0=-\infty$。这样，便可以得到如下寻求方向*w*和参数*b*、ξ、ξ^*的最优化问题。

$$\min_{w,b,\xi,\xi^*}\quad \frac{1}{2}\|w\|^2+C\sum_{j=1}^{2}\Big(\sum_{k=1}^{j}\sum_{i=1}^{l_k}\xi_{ki}^j+\sum_{k=j+1}^{3}\sum_{i=1}^{l_k}\xi_{ki}^{*j}\Big) \tag{1-12}$$

$$s.t.\quad ((w\cdot x_i^k)-b_j)\le -1+\xi_{ki}^j,\quad \xi_{ki}^j\ge 0,\quad k=1,...,j, i=1,...,l^k; \tag{1-13}$$

$$((w\cdot x_i^k)-b_j)\ge +1-\xi_{ki}^{*j},\quad \xi_{ki}^{*j}\ge 0,\quad k=j+1,...,3, i=1,...,l^k; \tag{1-14}$$

其中*j*=1,2。

类似地，通常该问题的解也是通过求解其对偶问题来获得，并且经典的SMO算法可以应用到其对偶问题。这便于支持向量顺序回归机模型可以求解大规模问题。更为具体的讨论可以参见文[48, 49]。

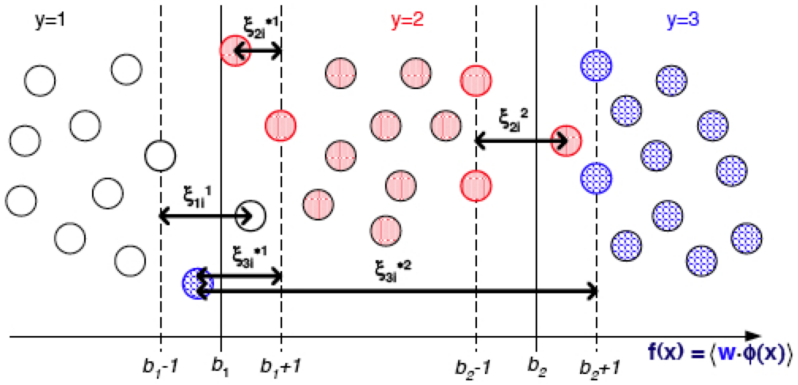


图1-5 支持向量顺序回归机模型图示

（摘自文[47]）

1.4 研究问题背景和前人工作介绍

蛋白质分子与其他蛋白质分子或者核酸大分子之间的相互作用是分子生物学和系统生物学研究的一个重点和难点。生物体内细胞的许多重要生理活动过程诸如信号转导、细胞周期调控、癌症疾病发生等都是通过蛋白质与其他大分子相互作用实现的。本节主要介绍蛋白质与其他大分子相互作用的背景知识和前人在这些问题上的相关工作。

1.4.1 蛋白质与核酸分子相互作用研究

在研究具体的蛋白质与核酸相互作用之前，需要首先确定出DNA结合蛋白质和RNA结合蛋白质。DNA结合蛋白质和RNA结合蛋白质是细胞体重要的功能蛋白质。DNA结合蛋白质在许多生物过程诸如DNA缠绕（packaging），DNA复制，以及控制基因表达等方面起着关键作用。而RNA结合蛋白质则通过在蛋白质合成的各个不同的生理阶段与RNA发生相互作用以实现其控制蛋白质合成进程的作用。在过去的十几年中，生物学家们用实验的手段测定了许多DNA结合蛋白质和RNA结合蛋白质。但是目前仍有大量的核酸结合蛋白质并未被实验完全测定，并且通过生物实验对蛋白质功能进行一一注释仍然非常昂贵并且费时。因此，通过生物信息学的技术手段预测未知的DNA结合蛋白质与RNA结合蛋白质受到了越来越多的重视和应用。

在前人的工作中，有很多基于进化信息[50]和蛋白质结构信息[51]的预测模型。也有一大类是基于机器学习方法构建的预测模型。这其中比较典型的工作有：2003年蔡煜东等人[52]利用蛋白质的伪氨基酸组成结合支持向量机的方法实现了对DNA, rRNA, RNA结合蛋白质的预测；2006年李亦学等人则在文[52]的基础上，从蛋白质的序列信息角度出发，利用伪氨基酸组成、物理化学性质等特征结合支持向量机方法实现了DNA/RNA/rRNA结合蛋白质的预测，分别取得了～72%、78%和84%的预测精度[53]。此外，Han等人于2004年在文[54]中提出利用蛋白质的一级序列信息，结合支持向量机的机器学习方法来实现对RNA（rRNA/mRNA/tRNA）结合蛋白质的识别预测。

在上述预测模型的构建过程中，很重要的一点是如何描述蛋白质，也即对蛋白质序列实现数值化编码。不同的特征编码方式对应着不同的生物学信息，如何有效而精准的描述出某类功能蛋白质（如DNA结合蛋白质或者RNA结合蛋白质）所特有的特征，对于预测模型的构建具有相当关

键的作用。在这个方面，已经有很多学者尝试过从蛋白质的不同信息角度进行特征编码，如蔡等人提出的基于蛋白质序列的伪氨基酸组成信息[52]；Ahmad等人整合的结构信息[55]和Nitin Bhardwaj等学者提出的基于生物化学特性的信息[56]等。但是如何更为精准而有效地对关心的功能蛋白质进行数值化编码仍将是非常有趣而重要的方面。同时，已有很多学者在研究其他同类问题的时候，提出了许多描述蛋白质的数值编码方式。对于这些新的特征编码方式是否能有效的用来描述核酸结合蛋白质，将仍是一个值得研究的重要课题。

1.4.2 蛋白质与蛋白质分子相互作用研究

蛋白质与蛋白质相互作用研究已经成为系统生物学界研究的一个热点问题，通过前面提及的高通量实验手段，人们已经得到了部分蛋白质与蛋白质相互作用数据。但是这部分实验测得的相互作用数据，相对于生物体内真正可能存在的相互作用对而言，显得微不足道。同时，各个不同的实验手段都存在着不同程度的“假阳性”和“假阴性”，得到的相互作用数据之间的重合度非常的低，这说明实验手段都有着各自的缺陷和不足。为了尽可能多确定出生物体内的蛋白质与蛋白质相互作用对，并且可以有效的处理上述实验存在的缺点，研究人员开始寻求从计算模型角度来进行预测蛋白质-蛋白质分子相互作用的工作。目前，相关的工作主要有：从基因组信息出发的种系发生谱[57]、基因融合[58]、邻域基因法[59]等；从蛋白质结构域（Domain）的角度出发的方法：关联方法（从已有的蛋白质-蛋白质互作数据中计算出比较频繁出现的结构域对[60]）、最大似然估计法[61]、结构域对排除分析[62]、最小*p*-值法[63]及Parsimony法[64]等；以及直接从蛋白质序列角度出发利用机器学习方法进行预测蛋白质-蛋白质相互作用的方法：基于序列信息（Domain、GO信息等）的支持向量机方法[65]、随机决策森林方法[66]等。对于利用机器学习的前述方法，大多都集中在从蛋白质序列角度来提取蛋白信息，或者结合了GO功能注释的信息或者结构域（Domain）等其他信息，但预测精度多集中在70%左右，没能取得较大的改进。2007年，蒋华良等人[67]提出了一种新的描述蛋白信息的新方法——“三联体”法，并结合采取新的对称核的支持向量机方法，使得其预测精度达到了84%左右，这对从序列层次进行蛋白质-蛋白质相互作用预测提供了新的契机。而这种新的描述蛋白质特征的方法也为其他功能蛋白质的识别提供了借鉴。

在过去一段时间的研究中，研究人员发现许多蛋白质相互作用是通过相互作用结构域（如SH2、SH3、WW, PDZ结构域等）与其识别的另一个蛋白质中一段短肽序列（大约10个氨基酸长度）的结合完成的，并且这些识别的短肽通常具有一些比较特定的特点（短肽识别motif），如SH3识别PxxP模体（motif），WW识别PPxY模体，而PDZ则识别C-末端疏水性氨基酸等。蛋白质相互作用结构域在蛋白质亚细胞定位、信号传导和蛋白复合体的组装等方面发挥重要功能。对于一个蛋白质的某一特定结构域，期望能在全蛋白质组的规模上完成对其结合特异性的研究，这将有助于全面系统地了解特定结构域的特点和功能，以整合到蛋白质相互作用网络中，解释特定的生物学机制。

本文另一部分工作正是集中在蛋白质结构域与短肽相互作用方面的研究，主要围绕WW结构域和PDZ结构域两个常见的信号传导结构域进行研究。在介绍这两个结构域的相关研究工作背景之前，有必要先介绍一下其他一些结构域-短肽相互作用领域的相关工作。在该领域，目前研究

工作者们已经取得了一些卓有成效的工作。这部分工作主要从结构生物学、计算结构生物学和生物信息学的角度开展研究工作[68-70]。

通过实验的手段获取结构域–短肽的相互作用复合物是此种类型数据获取的一个重要途径，也是分析复合物中物理接触结合位点和功能分析的一个基本途径。但是该方法受到实验条件的限制，仅局限于对单个结构复合体进行分析，无法应用到大规模的结构数据中，这也就很难从更高的层面分析出较有意义的结果。

计算结构学家通常通过收集现有的带有结构信息的结构域-短肽相互作用对数据，构建关于蛋白结构域-短肽复合物的能量函数模型。Fernandez-Ballester等人[71]通过构建三维结构模本进行短肽配体序列的同源建模并对每一个SH3结构域构建基于短肽序列位置特异的结合矩阵，以此预测SH3结构域的短肽结合特异性。Hou等人[72]采用了非常相似的基于三维结构的模型并结合机器学习的模型实现了对SH3结构域的短肽结合特异性的预测。不同之处是Hou等人的工作整合了所有SH3结构域家族的数据构建了一个统一的预测模型。类似地，Ferraro等人[73]基于已知的三维结构数据，将SH3结构域-短肽相互作用复合物中的氨基酸物理结合对作为模本信息扩展到其他未知结构信息的SH3结构域-短肽对中，并构建了机器学习模型。Zhang等人[74]则基于结构信息考虑氨基酸的生物化学特性，结合神经网络和支持向量机模型构建了针对SH2和PDZ结构域的结合特异性的预测模型等。此外，Wunderlich等人[75]根据现有SH2结构域的结构信息为模版构建了一个新的能量函数，以此预测SH2结构域-短肽相互作用。该模型不仅从结构角度找到了SH2结构域-短肽相互作用对的物理接触对，还结合信息理论获取了“协同进化”（co-evolution）的物理接触对。针对这部分物理接触对的建模分析，可以更好的理解影响SH2结构域–短肽相互作用的关键结合位点。

近来，越来越多基于生物信息学方法的预测模型被提出用于预测结构域-短肽之间的相互作用，并且取得了显著的效果。其中大部分是基于位置特异性打分矩阵（position-specific scoring matrices, PSSM）模型的工作，如Lehrach等人[76]、Obenauer等人[77]、Reiss[78]和Wiedemann等人[79]的工作等。也有基于序列信息结合机器学习方法而提出的模型。例如，McLaughlin等人在2006年提出的基于隐马尔科夫（Hidden Markov Model, HMM）模型的预测SH2结构域–短肽相互作用的工作[80]等。与计算结构生物学方法相比，基于生物信息学方法的工作更多的是基于大量的数据样本信息。而计算结构生物学方法，大部分仍是基于现有结构数据，很难推广到其他无结构的数据，同时计算量非常巨大。

此外，值得一提的是“主要组织相容性复合体”（Major Histocompatibility Complex，MHC）结构域-短肽相互作用的相关工作。MHC是人类体内一类非常重要的免疫蛋白结构域家族。关于MHC结构域-短肽相互作用的计算模型方面的工作已经较为丰富，有针对“单个MHC结构域”的PSSM类模型[81]，也有针对“多个MHC结构域”或者“MHC结构域家族”的诸多机器学习模型[82]。有针对定性预测MHC结构域-短肽是否相互作用的分类模型[83, 84]，也有定量预测MHC结构域-短肽相互作用强弱的回归模型[85, 86]。虽然MHC属于免疫蛋白结构域，不是传统意义上的信号传导结构域家族，但是基于其的结构域-短肽相互作用的工作给从事信号传导结构域WW或者PDZ结构域的相关研究提供了非常重要的参考价值。

1.4.2.1 WW结构域与短肽相互作用研究

WW结构域是一种常见的信号传导结构域。含有WW结构域的蛋白质通常是通过与含有丰富脯氨酸（Proline）的短肽序列的蛋白质相互作用来实现其生物学功能。目前，针对WW结构域与短肽相互作用的研究主要集中在WW结构域的结合特异性问题上，并且相关工作开展的还不是很多。

Otte等人最初于2003年利用NMR技术结合短肽scanning技术测得了42个WW结构域的短肽相互作用特性[87]。2004年Hu等人针对人类WW结构域，通过实验的方法获得了人类WW结构域与短肽相互作用的网络图谱。并针对较强的相互作用对，确定了部分WW结构域的结合特异性[88]。之后，Ingham等人于2005年利用蛋白质谱技术测定了人类中10个WW结构域的短肽相互作用特性[89]。2006年Hesselberth等人利用蛋白质芯片技术测定了酿酒酵母中WW结构域与其他蛋白质之间的相互作用谱图[90]。

上述这些工作，基本都是围绕着实验手段来展开，既费时又费金钱，而且很多实验并未真正实现高通量技术。随着研究的深入和部分高通量数据的获得，采用计算的办法来进行预测WW结构域的结合特异性变得极为重要和紧迫。但是目前基于计算模型预测WW结构域-短肽相互作用的工作仅有Hu等人提出的利用PSSM方法进行短肽配体预测的工作和Wade小组采用结构模拟和同源对比的方法实现WW结构域-短肽相互作用预测的工作[91]。这些方法都是基于单个WW结构域而设计的，很难推广到整个WW结构域家族。

此外，测定结构域-短肽相互作用的亲和力大小（或者相互作用强弱的不同水平）也是一件极为重要的事情。而上述计算模型都没有考虑这一点。尝试在该问题上设计行之有效的计算模型将给生物学家带来更为便利而有用的信息。

1.4.2.2 PDZ结构域与短肽相互作用研究

PDZ结构域是另一种常见的信号传导结构域。含有PDZ结构域的蛋白质通常通过识别其他蛋白质的C端短肽序列来实现其生物学功能。PDZ结构域与短肽相互作用一般根据P_{-2}的氨基酸的类型被定义为三类：第I类PDZ结构域识别 Ser/Thr-X-ψ-COOH，第II类识别 ψ-X-ψ-COOH，和第III类识别 Asp/Glu-X-ψ-COOH。其中ψ为亲水氨基酸，X为任一氨基酸。

目前，针对PDZ结构域与短肽相互作用的研究主要从两个方面展开。第一类是关心PDZ结构域的结合特异性，第二类是研究PDZ结构域与短肽相互作用的亲和力（相互作用强弱）。

在结合特异性方面，除了上述三类传统的结合类型外，随着研究的深入，人们对这种传统的划分PDZ结构域的方式越来越有不同的争议。传统的实验手段，往往是针对单个PDZ结构域或者少数几个PDZ结构域而开展的，得到的PDZ结构域的结合特异性的信息就会在很大程度上受到局限。为此，Tonikin等人于2008年对Human和Worm物种中85个PDZ结构域进行了噬菌体展示的高通量生物实验，通过实验得到PDZ结构域根据结合特异性可以分为16类[92]。这一针对PDZ结构域家族的结合特异性图谱的研究工作更为系统地揭示了PDZ结构域的生物功能特征，并且为进一步发现新的PDZ结构域提供了有效的信息资源。此后，在此实验数据的基础上，有很多学者开展了预测给定PDZ结构域和短肽序列，它们是否发生相互作用的工作，如Hui等人的工作[93]。但是这些相关的工作还远远没有达到可以实现针对各种不同的物种中的PDZ结构域和各自生物体内的真实短肽之间的相互作用预测的要求。进一步整合不同的数据资源，开发预测模型，将是该领

域未来几年致力于研究的一个重要方向。

在PDZ结构域-短肽相互作用亲和力方面，相关的研究还很少。其中第一个高通量测得PDZ结构域-短肽相互作用强弱的数据集是Stiffler等人在2007年发表在《Science》上的关于小鼠PDZ结构域的数据[94]。之后同一个实验室的Chen等人于2008年更新了Stiffler的实验，并在线虫和果蝇物种中得到了部分PDZ结构域-短肽相互作用的数据[95]。同时，两者都基于此数据集设计了相关的计算模型，为进行下一步实验提供了指导信息。值得注意的是，这些模型主要都是用以预测PDZ结构域-短肽对之间是否发生相互作用，并未真正的涉及到预测PDZ结构域-短肽相互作用的强弱。但是鉴于知道PDZ结构域-短肽相互作用的强弱对于认识生物过程的重要性，设计和开发合适的计算模型显得迫在眉睫。

1.5 论文研究的问题和组织结构

本论文主要利用机器学习的方法特别是基于支持向量机的方法对系统生物学中蛋白质与核酸、蛋白质与蛋白质之间相互作用的若干预测问题进行了探索、研究。通过对生物问题的深入理解，将生物问题中涉及的预测问题抽象成为数学问题，并建立相应的模型加以求解。论文的主要内容安排如图1-6。论文首先关注蛋白质与核酸分子相互作用的生物学问题，第二章重点针对DNA结合蛋白质和RNA结合蛋白质的识别预测开展研究。这两类核酸结合蛋白质在细胞体内许多生物过程中起着非常重要的作用，实现对这类功能蛋白质的自动预测非常有意义。接着开展关于蛋白质与蛋白质相互作用中一类重要的蛋白质结构域与短肽相互作用的研究。这一类短肽识别结构域与其他普通的结构域不同，它们有着具有高度结合特异性的识别模体（motifs）。结构域–短肽相互作用预测问题主要包括预测它们与短肽配体之间的结合特异性和结合亲和力。较为理想的状况是知道结构域与短肽相互作用的亲和力，但是目前碍于实验技术的限制，并不能对所有的短肽识别结构域开展获取亲和力数据的实验。论文第三章和第四章属于同一个研究范畴，分别关注WW结构域和PDZ结构域与其短肽配体的相互作用问题。这两个结构域都是在信号传导、蛋白质复合物装配过程中起关键作用的结构域，与人类的许多疾病有关。由于针对这两个结构域的研究采用的实验手段不同，获得的数据类型不同，本文采取的研究策略也略有不同。论文第三章重点考虑预测WW结构域–短肽是否发生相互作用以及相互作用强弱水平的问题。而第四章则重点考虑定量化预测PDZ结构域与短肽相互作用亲和力的问题。

论文的具体内容安排如下：

第一章，简要介绍分子生物学的基本概念、系统生物学的背景知识以及相关的生物学数据库。同时，还简要介绍了本文涉及到的若干支持向量机模型以及本文研究的生物学问题的背景知识，指明了全文的主要研究内容与框架。

第二章，介绍本文基于蛋白质氨基酸序列信息预测DNA/RNA结合蛋白质的工作。对现有的测定DNA/RNA结合蛋白质的实验手段和计算预测方法进行了简单的综述。将对蛋白质序列进行有效描述的“三联体”特征表示方法应用到DNA/RNA结合蛋白质的预测问题中，构建了基于支持向量分类机的预测模型，并通过数值试验验证了方法的有效性。

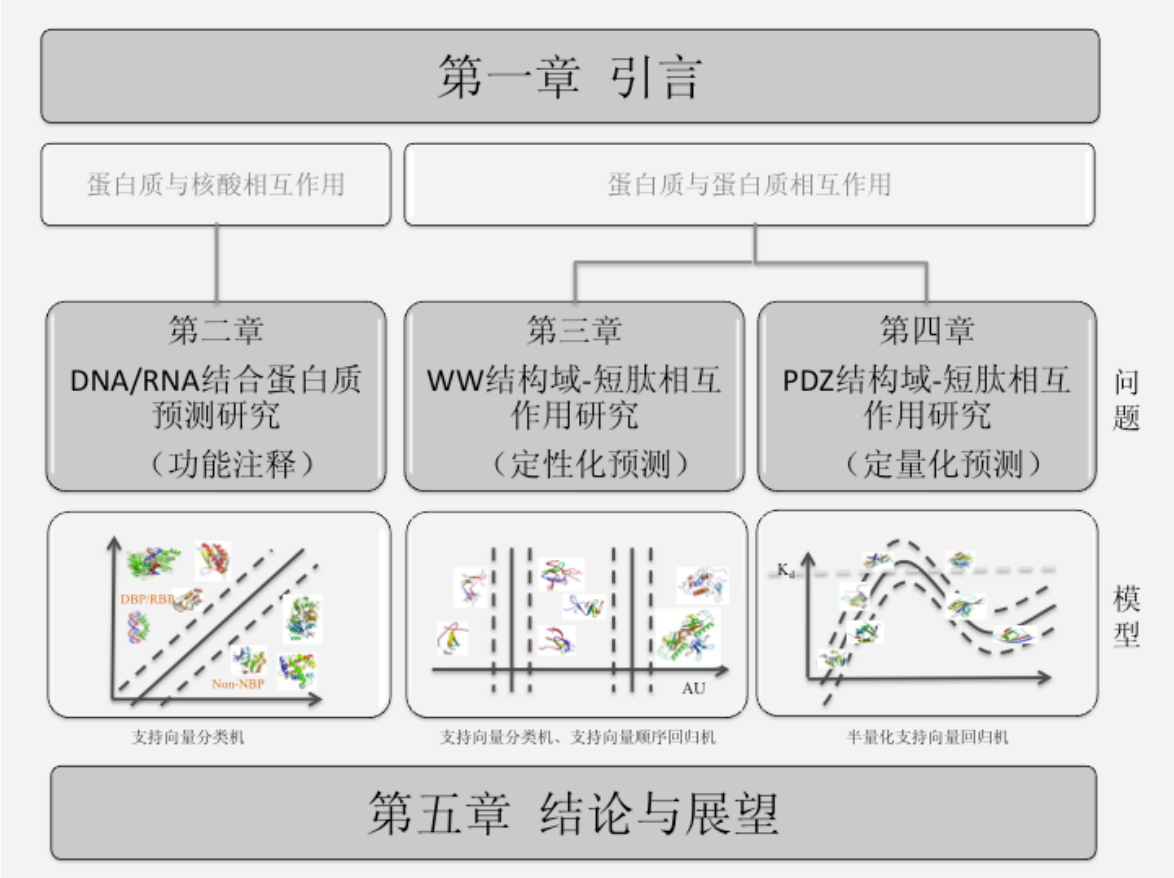


图1-6 本论文主要内容安排

第三章，简单介绍WW结构域–短肽相互作用的实验方法和计算方法，提出了基于支持向量分类机的预测WW结构域-短肽相互作用的两分类模型。此外，根据生物学家的不同需求，设计了基于支持向量顺序回归机的预测WW结构域-短肽相互作用对的强弱水平的预测模型，为进一步研究WW结构域的结合特异性提供便利框架。

第四章，根据Stiffler和Chen等人实验测定的关于PDZ结构域-短肽相互作用的具有亲和力大小的数据特性，建立了相应的回归预测模型。该数据集同时包含了具有亲和力大小的定量化数据（亲和力大小<100uM），也包含了只含负类样本信息的定性化数据（亲和力大小≥100uM）。针对该数据集的特性，本文设计了同时考虑定量化的正类样本和定性化的负类样本的支持向量回归模型——半量化支持向量回归机。本文的工作实现基于小鼠内PDZ结构域-短肽相互作用数据的对PDZ结构域-短肽相互作用亲和力大小的预测。

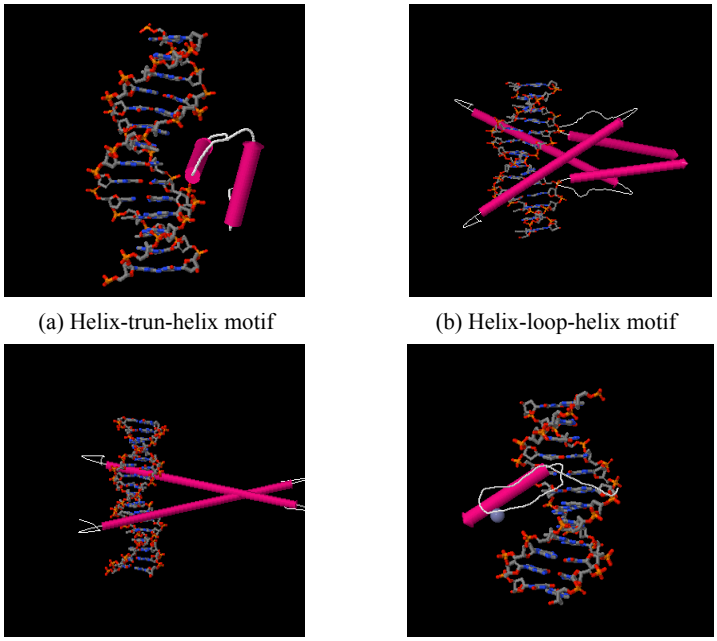
第五章是对论文工作的一个总结以及对下一步工作的建议。

第二章 DNA/RNA结合蛋白质预测研究

DNA结合蛋白质和RNA结合蛋白质是细胞体内重要的功能蛋白质。确定DNA结合蛋白质或者RNA结合蛋白质可以帮助人们理解生物体内基因的转录调控过程和翻译过程等。通过实验手段，生物学家已经注释了大量的DNA结合蛋白质和RNA结合蛋白质，但这一类核酸结合蛋白质的功能注释过程仍然远未完成。由于实验的手段既费时又费力，人们开始从计算预测模型的角度去注释DNA结合蛋白质或者RNA结合蛋白质，但是模型精度仍有很大的改进空间。本章的工作是对前人工作的延续和改进，采用新的描述蛋白序列的特征编码方式，构建了基于支持向量分类的预测DNA/RNA结合蛋白质的分类器。

2.1 引言

DNA结合蛋白质主要是由DNA结合结构域构成，并且与单链或者双链DNA发生相互作用。该蛋白质通常与单键DNA相互结合的亲和力极大，与双链DNA结合的亲和力较差。它们在诸多的生物过程中起到重要的作用，比如DNA转录过程中起到调控转录功能，具有DNA缠绕（packaging）功能，控制DNA复制以及控制基因表达等功能。相应地，DNA结合蛋白质主要包括转录因子（Transcription factors），各种不同的聚合酶，以及在染色体缠绕过程中起作用的组蛋白（Histone）等，其中转录因子是最主要也是最重要的一类DNA结合蛋白质。DNA结合蛋白质通常与DNA的大groove部位发生相互作用，也有少部分DNA结合蛋白与DNA结合会发生在DNA的小groove的部位[96]。根据DNA结合蛋白质与DNA相互作用的模式不同，可以将DNA结合蛋白质进行分类。目前，DNA结合蛋白质主要有四大类[97]：螺旋-转角-螺旋（helix-turn-helix）结合模体（Motif）、螺旋-卷曲-螺旋（helix-loop-helix） 模体（bHLH）、亮氨酸拉链（leucine zipper）模体（bZIP）、锌手指（zinc finger）模体等，见图2-1。此外，DNA结合蛋白质还包括了识别TATA盒子的TBP结合结构域等。RNA结合蛋白质主要与单链或者双链的RNA发生相互作用以此来调节RNA的翻译、控制翻译后修饰如RNA剪切、编辑等事件。RNA结合蛋白质通常是细胞质和核蛋白，包括翻译引导蛋白、polyA 结合蛋白、snRNPs和ADAR等。类似与DNA结合蛋白质，通常根据RNA结合蛋白质与RNA结合的模式不同将其进行分类。由于RNA是一种结构化的分子物质，RNA结合蛋白质所具有的模体（motif）也相对丰富一些，目前发现的主要有[98]：RNA-识别模体（RNA-recognition motif）、K-同源(K-homology，KH)结构域、双链RBD（double-stranded RBD）、RNA结合锌手指（Zinc fingers） 模体等，见图2-2。



(a) Helix-trun-helix motif (b) Helix-loop-helix motif

(c) Leucine zipper motif (d) Zinc finger motif

图2-1 DNA结合蛋白质模体图示

(a) 螺旋-转角-螺旋模体；(b) 螺旋-卷曲-螺旋模体；(c) 亮氨酸拉链模体；(d) 锌手指模体。图中双螺旋状为DNA，另一半为DNA结合蛋白质。上述4幅图均摘自如下网页：http://arapaho.nsuok.edu/~biology/Tutorials/DNAbinding.htm.

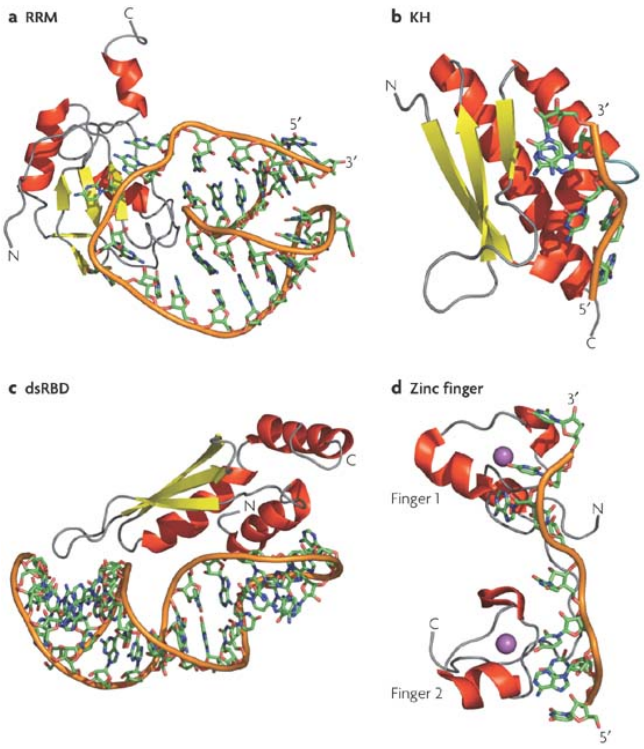


图2-2 RNA结合蛋白质模体图示

(a)RNA识别模块；(b)KH结构域模体；(c)双链RBD模体；(d)RNA结合锌手指模块。上述4幅图均摘自文献[98]。

虽然随着人类基因组计划的完成，越来越多的基因及其产物（蛋白质）得到了正确的注释，但是在当前公开数据集中仍然有大半数以上的蛋白质没有得到其相关功能的注释。完成对未知功能蛋白质的注释是生物学家一直致力于研究的事情。其中DNA结合蛋白质和RNA结合蛋白质就是两类需要进一步注释的重要功能蛋白质。除了前述所介绍的它们在生物过程起的重要作用之外，近来很多文献表明DNA结合蛋白质或者RNA结合蛋白质与人类很多疾病如癌症等有关。正是由于DNA结合蛋白和RNA结合蛋白在生物过程中和相关的疾病研究中起着如此重要的作用，对DNA/RNA结合蛋白的注释一直是蛋白质组领域内一个比较热门而重要的研究课题。

在过去的十几年中，生物学家们用实验的手段测定了许多DNA结合蛋白质和RNA结合蛋白质。目前常用的测定DNA/RNA结合蛋白质的实验技术主要有以下几种[99]： 酵母单杂交、 DNA-蛋白质印迹杂交、 凝胶迁移或电泳迁移技术（EMSA）、 DNaseI足迹法、蛋白质芯片技术、染色体免疫沉淀技术（Chromatin Immunoprecipitation，ChIP）和ChIP-chip技术等。此外，还有DNA与蛋白质复合体的电镜观察、干扰实验、紫外交联法等方法。

虽然生物实验的手段在最近几年已经得到了长足的发展，但是要通过实验的手段测定所有未知的DNA/RNA结合蛋白质复合物已然不合时宜。近年来，生物信息的技术手段受到了越来越多的重视和应用。

在生物信息学中，对于蛋白质与DNA/RNA等核酸分子相互作用的研究，国内外学者已经开展了一系列相关的工作，并取得了许多卓有成效的成果。这其中包括经典的利用序列比对的方法[100]、基于进化信息的方法[50]、结构基因[51]和近邻基因[59]的方法等。虽然大多数的蛋白质功能注释是基于蛋白质序列的相似性来进行，但是一半以上的蛋白质却没有非常相似的已知功能的蛋白质可以作为参考。DNA结合蛋白质是其中一类比较大的蛋白质家族，它们在蛋白质序列层面或者结构上各异。在结构层面上，它们可以大致分成54大类[101]。正是因为DNA结合蛋白质之间具有如此大的差异性，使得传统的基于序列相似性的办法，进化树的办法来进行功能的注释往往得不到较好的结果。近年来有学者从另外的角度来考虑预测问题，其中一类比较成功的办法是基于机器学习模型的方法。具体地，主要通过机器学习模型预测是否是核酸配体结合蛋白质的工作，进而确认相关的相互作用网络，如基于决策树[102]、神经网络[103]、支持向量机[52]的方法等。在所有机器学习的方法中，很重要的一点是如果将关心的蛋白质数据进行数值化编码，每种不同的编码方式蕴含了不同的信息。

基于机器学习方法的工作中根据编码方式的不同，大体上可以分为两大类，第一类是基于蛋白质结构的工作，第二类是基于蛋白质序列的工作。在基于结构信息的工作中，主要有：Stawiski等人[104]于2003年发表的工作，他们通过蛋白质结构信息来计算蛋白表面正电荷处的静电斑块（patch）等特征信息，并结合神经网络来进行预测DNA结合蛋白质。Shanahan等人[105]则利用DNA结合蛋白质复合物之结构化模体和静电量信息来进行计算预测DNA结合蛋白质。2004年Ahmad等人[103]整合了蛋白质的序列组成、蛋白质的可溶性和二级结构信息等和神经网络方法实现了对DNA结合蛋白的识别。同年，Ahmad等人[55]还利用净电荷信息、偶级矩和四极矩的信息实现了对DNA结合蛋白质的预测。2005年Nitin Bhardwaj等学者[56]利用总电荷、静电量、斑块（patch）、氨基酸组成来构建蛋白质的特征，再利用支持向量机进行分类器的设计，最终实现对DNA结合蛋白的预测。Shazman等人于2008年提出了描述RNA结合蛋白的基于三维结构信息的新

特征表示方式，并利用支持向量机对RNA结合蛋白识别问题构建了分类器，得到了较高的分类性能，分类精度高达88%[106]。上述工作均达到了较高的分类性能。但是上述方法因为是基于结构信息（而这种结构信息往往是很难获得）的，并不能进行大规模的注释，大大限制了这种类型的模型方法的应用。

在基于蛋白质序列信息的工作中，如第一章所介绍的部分内容，主要有2003年蔡煜东等人利用蛋白质的伪氨基酸组成结合支持向量机的方法[52]、2006年Yu等人在文[52]的基础上利用伪氨基酸组成、物理化学性质等特征结合支持向量机方法实现了对DNA/rRNA/RNA结合蛋白质的预测[53]。而Han等人则利用蛋白质的一级序列信息，结合支持向量机的机器学习方法实现了对RNA（rRNA/mRNA/tRNA）结合蛋白的识别预测[54]。前人的工作为有效地进行DNA/RNA结合蛋白质的预测奠定了基础，但是从计算的角度而言，仍有进一步提升预测精度的空间，同时新的对蛋白质编码的方法也陆续被提出，尝试新的编码方式来进行DNA/RNA结合蛋白质的预测非常有意义，可以从不同的角度揭示出具体的生化机制。

本章结合了一种在蛋白质-蛋白质相互作用预测研究中提出的用来描述蛋白质的“三联体”的特征编码方式[67]，利用支持向量机分别构建分类器对DNA/RNA结合蛋白质进行预测。通过数值试验发现“三联体”的特征编码更为有效地提供了识别DNA结合蛋白质的信息。同时对于RNA结合蛋白质，“三联体”的特征编码也能有效地表示。此外，本章还利用“三联体”的特征编码，结合SVM实现了对DNA结合蛋白质和RNA结合蛋白质的分类。通过特征选择，我们得到了对于各个不同的分类器最富含分类信息的“三联体”特征，并且在部分已知蛋白质结构的复合体中发现了这些“三联体”特征往往处于蛋白质接触表面，这表明“三联体”特征在一定程度上很好地抓住了蛋白质结合位点的特征。

2.2 数据集和方法

2.2.1 数据集

本章所采用的数据集主要从Swiss-prot数据库（第52个版本）[107]上下载。通过关键词“DNA-binding”和“RNA-binding”等获取得了所有DNA/RNA结合蛋白质，这些蛋白质将作为分类问题的“正类”样本点。分类问题的负类样本点的获取步骤主要有[108]：首先，通过搜索与DNA/RNA结合蛋白质相关的一系列关键词得到了一个“对照组”蛋白。其次，从整个Swiss-prot数据库中去掉上述“对照组”蛋白质后，得到了没有功能注释的蛋白质，即分类问题的“负类”点。

为了便于对蛋白质进行特征提取，在上述检索得到的蛋白质中，去掉了蛋白质序列长度超过6000个氨基酸或者少于50个氨基酸长度的蛋白质。同时，对于在蛋白质序列中含有特殊符号如“X”或者“Z”的蛋白质，也将其去掉。

通过上述处理得到的“正类”蛋白质和“负类”蛋白质之间存在着同源序列。为了消除同源序列带来的可能对分类器构建产生偏差的影响，本章采用常用的CD-HIT程序[109]，对上述蛋白质集合进行了去“冗余”处理，其中相应的同源性参数设为25%。至此，得到的“正类”和“负类”蛋白质之间的序列同源性均小于25%。

最终得到了1090个DNA结合蛋白质和358个RNA结合蛋白质，同时得到了16932个无功能注释的“负类”蛋白质。为了避免在进行支持向量机训练时数据不均衡所可能带来的问题，本章采用随机抽样的办法来平衡正负类的训练点。具体如下：对DNA结合蛋白质预测问题而言，在训练支持向量机的时候，从16932个“负类”点中随机选取1090个蛋白质作为训练所用的“负类”点；对RNA结合蛋白质而言，则相应地从中随机选取358个蛋白质作为训练所用的“负类”点。若无特殊说明，本章的预测结果均基于对该过程进行20次随机抽样之后得到的平均值。

此外，本章还试图建立可以区分DNA结合蛋白质和RNA结合蛋白质的分类器。在构建此分类器时，采用全部的1090个DNA结合蛋白质和358个RNA结合蛋白质作为训练样本点，分别定义为“正类”点和“负类”点。具体的蛋白质ID号可参见[108]。

为了考虑比较采用Swiss-prot数据库中的DNA结合蛋白质的全序列或者采用具有三维结构的PDB数据库[110]中相应的DNA结合蛋白质序列构建的分类器之间的分类性能，我们收集了具有三维结构的部分DNA结合蛋白质数据集和对应的具有三维结构的非结合蛋白质数据集，并从中选取了相应的蛋白质序列。这部分数据采集自文[111]，与该文中的“DNAset”集一致。在这部分数据集中，共收集了非冗余的DNA结合蛋白质146个(任意两个蛋白质之间的序列相似度不大于25%)，非冗余的非核酸结合蛋白质250个。为了方便起见，本章将从PDB数据库中收集的蛋白质序列称为蛋白质“结构域”数据，而将从Swiss-prot数据库中收集到的蛋白质序列称为全序列数据。

2.2.2 特征编码

利用机器学习方法进行蛋白质功能预测中一个关键问题是如何表示蛋白质的关键信息，也即特征编码的问题。本章采用一种新的蛋白质编码方式来表示蛋白质信息，称之为“三联体”特征编码[67]，该编码方式最初于研究蛋白质-蛋白质之间相互作用预测的工作中提出。如该文指出的，蛋白质-蛋白质之间的相互作用主要依赖于氨基酸的电荷（electronics）和疏水性（hydrophobic）的特性。可以想象蛋白质与DNA/RNA核酸分子之间的相互作用也很有可能是依赖于上述两种重要的生化特性。故此，本章尝试使用“三联体”编码，来考察其是否可以有效的表示DNA/RNA结合蛋白质。

该编码主要是考虑了各个氨基酸的侧链的偶极性（dipole）和体积（volume）的不同特性，并依据这两个特性将20个氨基酸归为7类，分别为{A, G, V}、{I, L, F, P}、{Y, M, T, S}、{H, N, Q, W}、{R, K}、{D, E}、{C}。具体参见表2-1。之后，在此基础上考虑将蛋白质序列中前后相连的三个氨基酸作为一个“单元”，通过此种“单元”的分布来描述整个蛋白质序列。

表2-1 20种氨基酸分类——分为七类

Class	Amino Acid(s)	Dipole scale	Volume scale ($Å^3$)
1	Ala, Gly, Val	<1.0	<50
2	Ile, Leu, Phe, Pro	<1.0	>50
3	Tyr, Met, Thr, Ser	1.0<Dipole<2.0	>50
4	His, Asn, Gln, Trp	2.0<Dipole<3.0	>50
5	Arg, Lys	>3.0	>50
6	Asp, Glu	>3.0(反向)	>50
7	Cys*	1.0<Dipole<2.0	>50

注*：将Cys归为区别于第3类的另外一类是因为其具有能形成二硫键的特性。

此外，考虑到蛋白质与DNA/RNA核酸相互作用时，往往是蛋白质序列的表面局部小片段与DNA或者RNA相互作用，所以设计“三联体”结构来构建其局部信息的特征描述方式是一个不错的选择。根据之前将20种氨基酸进行分类的结果，可以得到不同的“三联体”组合，进而可以得到描述整个蛋白质序列中不同的“三联体”分布的特征向量。关于蛋白质序列的三联体的编码过程可参见文[67]。注意，因为考虑了事先将20种氨基酸进行了归类，所以相同类之间对应的氨基酸构成的“三联体”，将视为具有类似的功能。例如三联体“AFM”与“GLS”将视为同一类“三联体”。同时，由于该编码方式将相似的氨基酸归为一类，大大缩小了特征空间的维度，使得机器学习得到的分类规则更具有一般性。

为了便于与其他基于序列信息的蛋白质编码方式进行比较，本章还利用了当前比较流行的Profeat工具对蛋白质进行编码[112]。Profeat是一个基于序列信息对蛋白质进行编码的有效工具，其共有七组特征，分别是氨基酸组成、双肽组成（Amino acid composition, Dipeptide composition）；规范化的Moreau-Broto自相关系数（Normalized Moreau-Broto autocorrelation）；Moran自相关系数（Moran autocorrelation）；Geary自相关系数（Geary autocorrelation）；蛋白质序列组成，氨基酸转换率及其分布（Composition, Transition, Distribution）；氨基酸序列顺序coupling统计、拟氨基酸序列序（Sequence-order-coupling number, Quasi-sequence-order descriptors）；以及伪氨基酸描述子（Pseudo amino acid descriptors）。其中第1组、第5组和第7组特征是其中应用最为广泛的对蛋白质序列进行编码的特征。该编码方式已经广泛应用于生物信息学中诸多涉及蛋白质功能识别的问题，如预测蛋白质-蛋白质相互作用，预测RNA结合蛋白质等。

2.2.3 分类器

采用支持向量机作为分类器，构建识别DNA/RNA结合蛋白质的分类模型。如第一章所述，采用经典的两分类支持向量分类机模型（C-SVC）。对应的DNA结合蛋白质或者RNA结合蛋白质，为各自分类器的正类点；而负类点则由如前面数据集中所述从无功能注释的蛋白质集合中随机抽取与“正类点”同规模的蛋白质构成。对应的输入向量是各自蛋白质的“三联体”编码之特征向量。

本章采用LibSVM作为实现支持向量分类机模型的软件，其中涉及到的参数利用网格（grid）搜索方法来搜寻最优参数[113]。

2.2.4 特征选择

本章采用基于互信息（Mutual Information）的最小冗余-最大相关性方法（Minimum Redundancy - Maximum Relevance, mRMR）[114, 115]进行特征选择。该方法已经成功应用于基因表达数据中的基因选择问题研究[114]。简言之，该方法可以选取对分类（如本研究中的“DNA结合蛋白质”和“非结合蛋白质”两类问题）最具有贡献的特征，并且可以实现各个特征之间具有最小的冗余程度，也就是说选择后得到的各个特征之间具有较大的不相似性。

给定第i个特征为 f_i 和相应的类别标号y, 那么根据对应的概率密度函数$p(f_i)$、$p(y)$和$p(f_i, y)$，它们之间的互信息可以定义为:

$$I(f_i,y)=\int p(f_i,y)\log\frac{p(f_i,y)}{p(f_i)p(y)}df_i\,dy\,. \tag{2-1}$$

为了量化各个特征对分类问题（如本研究中的“DNA结合蛋白质”和“非结合蛋白质”两类问题）的贡献度，使用最大相关性算法（MR），根据各个特征相应的互信息值$I(f_i, y)$的大小，选择最富含信息的前m个特征。也就是说，选择对涉及问题分类效果最佳的m个特征：

$$\max_S D=\frac{1}{|S|}\sum_{f_i\in S}I(f_i,y)\,, \tag{2-2}$$

其中S表示特征约简之后的特征集合。

至此，根据最大相关性算法得到最富含信息的前m个特征，但是越来越多的工作表明，因为最好的特征之间可能存在显著的相关性，所以“最好的m个特征的简单整合并非对应是最佳的m特征组合”[114]。为了最大限度的减少特征之间冗余性带来的问题，如文[114]所提出的，本章采用如下最小冗余标准：

$$\min_S R=\frac{1}{|S|^2}\sum_{f_i,f_j\in S}I(f_i,f_j)\,, \tag{2-3}$$

其中，上述式子考虑了任一一对特征组之间的互信息。最小冗余度标准将使得特征约简之后得到的特征组更具有代表性。

整合上述最大相关性算法和最小冗余度标准，可以得到如下最小冗余-最大相关性（mRMR）特征选择方法[114]。简言之，设F为全部特征集合，设当前得到的具有m-1个特征的特征集合对应为S_{m-1}, 那么下一步可以根据如下原则从剩余的特征集合$\{F - S_{m-1}\}$中选择的第m个特征：

$$\max_{f_j\in F-S_{m-1}}[D-R], \tag{2-4}$$

其中D和R是前述所提到的两个优化问题的目标函数（也即对应为两个标准）。

2.2.5 评价指标

对于一个分类器而言，通常需要用到的度量指标[116]有：

“真阳性”（true positive，TP）：正类数据中被预测为正类的数据；

“假阳性”（false positive，FP）：负类数据中被预测为正类的数据；

“真阴性”（true negative，TN）：负类数据中被预测为正类的数据；

“假阴性”（false negative，FN）：正类数据中被预测为负类的数据。

通过这几个指标，还可以得到：

灵敏度（Sensitivity）：Sen = TP/(TP+FN)，定义为原来是正类点被分类器正确预测为正类点的比例（或概率）；

特异性（Specificity）：Spec = TN/(TN+FP)，定义为原来是负类点被分类器正确预测为负类点的比例（或概率）；

精度（Accuracy）：Acc = (TP+TN)/(TP+TN+FP+FN)，即为被分类器正确划分的点占总样本点

（包括正类和负类）的比值。

此外，接受者操作特性曲线（receiver operating characteristic curve，简称ROC曲线）也是描述分类器性能的一个较为有效的评价标准，其曲线下面积（Area under curve，简称AUC）常常用来定量表示分类器的好坏。

2.3 实验结果分析

2.3.1 与已有工作的比较

本章对DNA结合蛋白质和RNA结合蛋白质分别建立了分类器，具体的模型构建过程如图2-3所示。首先根据前面提到的DNA结合蛋白质和RNA结合蛋白质与各自对应的等数目的“负类”蛋白组成分类器的数据集；之后将所有蛋白质序列进行“三联体”特征编码，并将得到的数据集作为SVM的输入，构建相应的SVM分类器。在测试阶段，同样首先对测试蛋白质序列进行“三联体”特征编码，然后利用之前得到的DNA结合蛋白质分类器或者RNA结合蛋白质分类器进行预测。

本章主要建立在Yu等人的工作基础上，对描述蛋白序列的特征编码进行了新的尝试，从而得到了不同的分类器。为了方便起见，本章得到的分类器因为采用了“三联体”特征，相应分类器记为triad-SVM，而Yu等人的分类器则简记为Seq-SVM。同时为了充分地比较这两种不同编码方式对应的分类器之间的异同，从以下几个测试方法展开：

1. 自检验测试（Self-consistency testing）。将训练集作为测试集，也即用训练集得到的分类器去预测所有用到的训练集。自检验测试主要是看分类器对当前已知的数据的预测能力。该检验的缺点是不能反映分类器的稳定性。

2. k折交叉验证测试（k-fold cross validation）。常用的有k=10折交叉验证(10-fold cross validation)和留一法交叉验证（Leave one out cross validation，或者称jackknife test）。所谓10折交叉验证，即将数据集均匀地分成十份，轮流将其中9份做训练，留出的1份做测试，10次的结果的均值作为对算法精度的估计，一般还需要进行多次10折交叉验证求均值，例如10次10折交叉验证，更精确一点。所谓留一法是一种特殊的k折交叉验证(k为训练样本点总个数)，每次轮流取出一个样本做测试，用其余的全部样本做训练。该检验方法被认为是能最为精确地反映出分类器的实际性能，同时该检验方法得到唯一的精度结果，从而该检验方法越来越受到大家的重视和使用。由于两者属于同类验证方法，而留一法又较为精确，所以本章主要采用留一法进行比较。

3. 随机测试集测试（Hold-out testing or bootstrapping testing）。该检验过程是一个有放回的随机抽样过程，每次随机抽取部分样本作为测试集，剩余的作为训练集，如此反复多次，取平均为分类器的最终精度的估计。重复次数越多，对分类器真实精度的逼近程度越高。与留一法测试不同，采用该测试方法每次得到的分类精度差异较大。

在同Yu等人的结果进行比较时，发现当采用自检验测试时，本章得到的DNA结合蛋白和RNA结合蛋白分类器的预测性能均比他们的结果好（表2-2），这表明“三联体”特征编码方式表示的蛋白数据之间具有较好的“自相容性”。而当采用留一法进行检验时，发现本章提出的新模型对

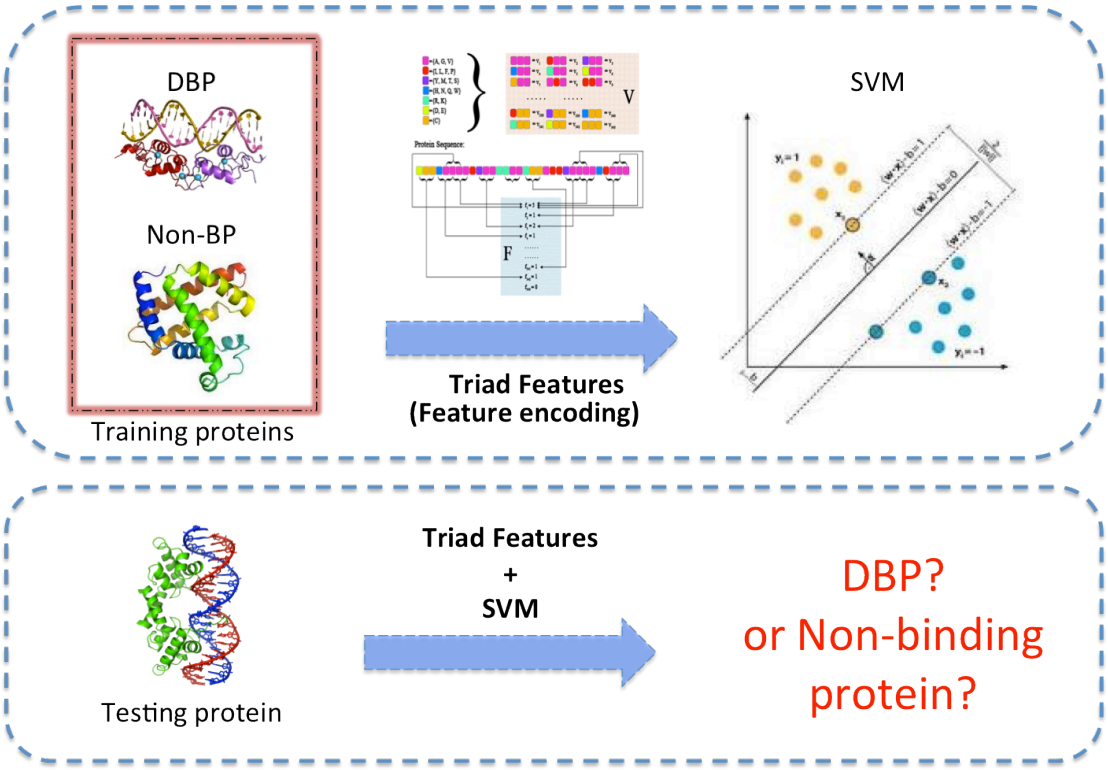


图2-3 DNA结合蛋白质分类器构建和预测图示

上部分为分类器训练过程，下部分为对未知新蛋白进行预测过程。此流程对RNA结合蛋白分类器同样适用。

DNA结合蛋白质具有更好的预测精度，预测准确率从之前的71.64%提高到了78.93%，并且无论从灵敏度还是特异性指标来看，均有所提高（表2-3）。而对于RNA结合蛋白质则得到了与Yu等人类似的结果，具体结果见表2-3。

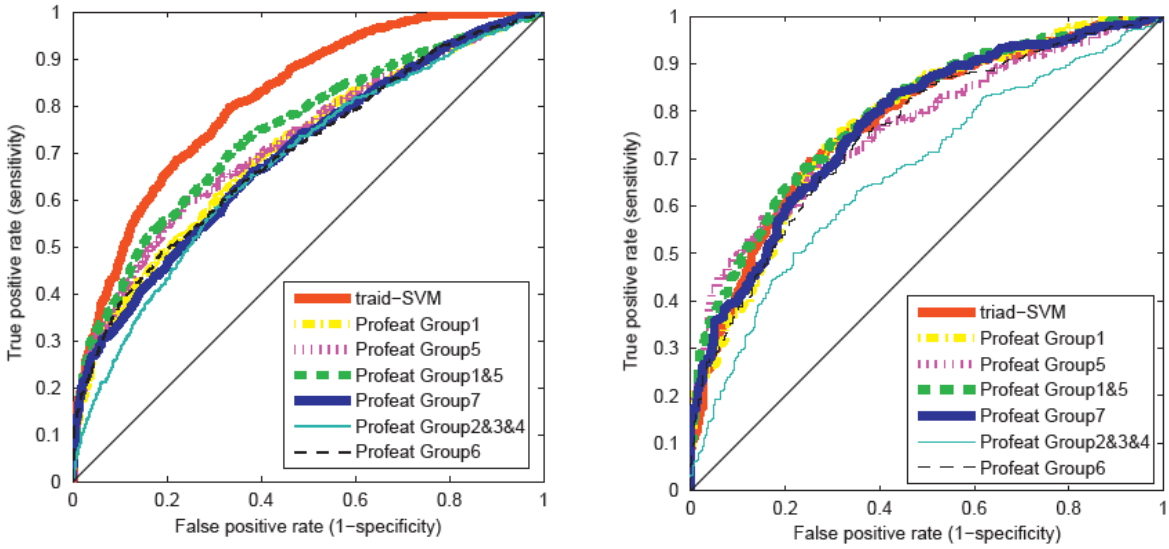
除了与Yu等人的结果进行比较外，本章也与更新的描述蛋白质序列特征的方法——Profeat进行比较。在与基于Profeat的特征编码的分类器（记为Profeat-SVM）比较时，采用留一法进行交

表2-2 基于自检验测试的triad-SVM和Seq-SVM结果比较

Proteins	Sample size	Methods	Accuracy(%)	Specificity(%)	Sensitivity(%)
DNA-binding	2180	Triad-SVM	90.37	73.75	91.40
		Seq-SVM	74.37	66.78	89.96
RNA-binding	716	Triad-SVM	89.37	90.24	89.55
		Seq-SVM	83.21	80.21	86.21

表2-3 基于留一法测试的triad-SVM和Seq-SVM结果比较

Proteins	Sample size	Methods	Accuracy(%)	Specificity(%)	Sensitivity(%)
DNA-binding	2180	Triad-SVM	78.93	66.74	84.86
		Seq-SVM	71.64	63.90	79.38
RNA-binding	716	Triad-SVM	76.75	74.81	78.70
		Seq-SVM	77.51	74.59	80.42



(A). DNA结合蛋白质分类器之ROC曲线 **(B). RNA结合蛋白质分类器之ROC曲线**

图2-4 DNA/RNA结合蛋白质分类器之ROC曲线图

叉验证测试，并以ROC曲线作为衡量分类器性能的指标。比较结果如图2-4所示。从该图中之A子图可以看出对于DNA结合蛋白质而言，Triad-SVM比基于Profeat的不同组的特征编码的所有分类的分类效果都好，具体地，Triad-SVM、Profeat-Group1、Profeat-Group5、Profeat-Group1&5、Profeat-Group7、Profeat-Group2&3&4和Profeat-Group6的平均AUC成绩分别为0.796、0.701、0.721、0.733、0.696、0.677和0.698。而对于RNA结合蛋白质而言，基于上述几种特征编码方式得到的分类器的预测精度几乎没有差别，对应的AUC值分别为0.776、0.773、0.767、0.792、0.775、0.684和0.754（图2-4B）。这进一步表明“三联体”特征对于DNA结合蛋白的识别更为有效。

此外，本章也采用了随机测试集方式[56]进行检验分类器的性能。如前所述，该测试方式得到的分类性能可以看作是真实独立测试集的预测性能。但是该测试方式的测试结果在一定程度上非常受如何划分训练集和测试集的影响，进而使得分类器的性能波动很大。本章利用box图方式展示各个分类器的分类性能，具体的结果参见图2-5。由图可见对于DNA结合蛋白质分类器，其预测精度从敏感度、特异性和总体精度三个方面而言，每次得到的精度都比较集中，显示了较为稳定的预测结果。相比而言，RNA结合蛋白质分类器在这些指标上则具有相对大一点的波动（图2-5）。其中DNA结合蛋白质分类器的平均分类性能为75.92%，不同的测试集对应的分类性能取值从71.33%到79.00%；特异性的平均性能为64.52%，灵敏度的平均值为86.31%。而RNA结合蛋白质分类器的平均性能为74.8%，分类性能变化从68.57%到80.00%不等；并且平均特异性取值为71.63%，平均灵敏度取值为77.96%。这也显示了“三联体”能更为有效地描述DNA结合蛋白质。

2.3.2 DNA-RNA结合蛋白质分类器

有文献表明DNA结合蛋白质与RNA结合蛋白质在很大程度上具有非常相似的特征[56]，很多RNA结合蛋白质甚至都可以与DNA结合（[117]）。本章尝试利用“三联体”特征编码对DNA结合

蛋白质和RNA结合蛋白质进行编码并构建支持向量分类模型，以察看是否可以从序列的角度进行有效的区分DNA结合蛋白质和RNA结合蛋白质。在该分类问题中，DNA结合蛋白质为1090个，RNA结合蛋白质为358个。利用三种不同交叉验证测试方式得到的分类预测性能之结果如表2-4所示。其中基于留一法的分类正确率达到了80.25%，该结果表明“三联体”特征编码能有效地区别出DNA结合蛋白质和RNA结合蛋白质。

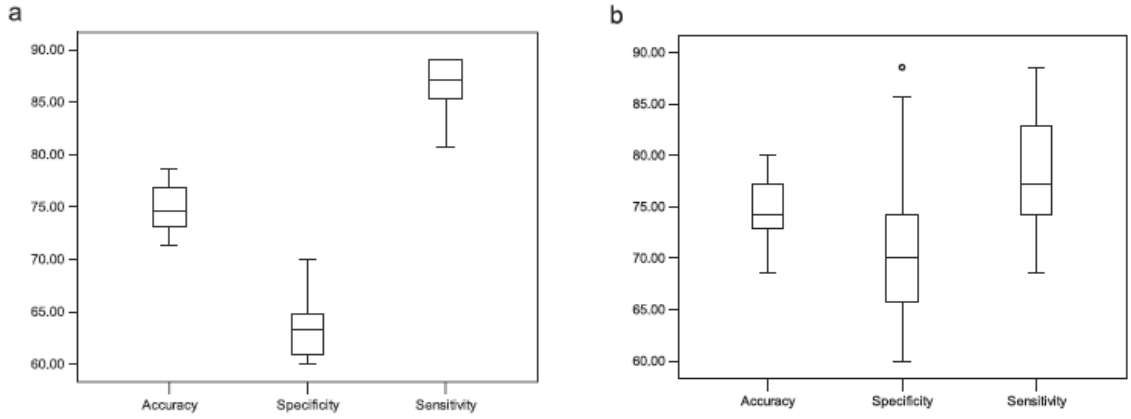


图2-5 两个分类器之随机测试集测试性能（Box图）

(a) DNA结合蛋白质分类器，(b) RNA结合蛋白质分类器。图中结果依次对应为分类精度，分类特异性和分类灵敏度。

表2-4 DNA结合蛋白质-RNA结合蛋白质分类器之分类结果

Test Methods	Accuracy(%)	Specificity(%)	Sensitivity(%)
Self-consistency	92.03	82.97	94.27
Jackknife	80.25	72.71	82.73
Hold-out	78.39	70.08	81.12

2.3.3 特征选择

此外，为了查看哪些特征起到关键的分类作用，本章还对上述“三联体”特征进行了特征选择。针对DNA结合蛋白质分类器、RNA结合蛋白质分类器和DNA-RNA结合蛋白质分类器，分别统计了各自重要的“三联体”特征。本章利用去冗余的mRMR特征选择方法，得到了针对上述三个分类器的前20个最富含分类信息的三联体特征，具体结果如表2-5所示。从表中可以看出DNA结合蛋白质分类器和RNA结合蛋白质分类器之间享有部分最具有分类特性的“三联体”特征，这表明DNA结合蛋白质与RNA结合蛋白质在某些特征上具有非常相似的特性。而从DNA-RNA结合蛋白质分类器的特征选择结果可以看出，大部分在DNA结合蛋白质分类器和RNA结合蛋白质分类器中享有的共同特征不再是区分DNA结合蛋白质和RNA结合蛋白质的主要特性，需要从其他“三联体”特征加以区别。

表2-5 前20个最具分类信息的“三联体”特征

Top Features	DNA-binding	RNA-binding	DNA-RNA
1	444	555	151
2	555	151	236
3	226	154	154
4	553	515	324
5	666	155	616
6	215	125	333
7	325	551	535
8	333	222	515
9	155	115	611
10	222	666	545
11	255	545	516
12	216	255	321
13	236	521	555
14	323	251	363
15	532	535	161
16	512	444	115
17	552	511	332
18	363	161	664
19	112	455	125
20	621	616	432

针对DNA结合蛋白质分类器、RNA结合蛋白质分类器和DNA-RNA结合蛋白质分类器利用mRMR特征选择方法进行特征选择。表中各数字中各个位置上的“1-7”代表“三联体”之各个位置上的氨基酸所归属的类别。

通过查看部分DNA结合蛋白质和RNA结合蛋白质复合物的三维结构，发现上述统计得到的“三联体”和DNA结合蛋白质与DNA发生结合或者RNA结合蛋白质与RNA发生结合的结合位点表面有关。例如，对于DNA结合蛋白质分类器, 三联体特征“444”，对应为连续的3个较高极性的大氨基酸（H, N, Q, W），例如1AZ0蛋白质中的A链中的三联体“QNH”和B链中的“QQN”以及“QNH”三联体等。其次为三联体特征“555”，对应为连续的3个碱性的大氨基酸(R, K)，例如PDB数据库中7GAT中对应的“KKR”三联体和9ANT对应的DNA结合蛋白质中的“RRR”三联体，具体图示可参见图2-6。其他的依次类同。

而对RNA结合蛋白分类器而言，三联体特征“555”也为最重要的分类特征；其次是“151”，对应为小亲水氨基酸（A, G, V）连着碱性的大氨基酸（R, K）再接着小亲水氨基酸（A, G, V）。例如PDB数据库中的1C0A中RNA结合蛋白中的三联体“RRR”；PDB结构数据1A34中RNA结合蛋白中的“VRA”，1ASY中的“GKA”就处于结合位点处等（注：采用5Å为阈值定义是否为结合表面，见文[118]）。其中部分图示可参见图2-6。其他的依次类同。

对于DNA-RNA结合蛋白质分类器，三联体特征“151”为最具有分类性能的特征，结合上面得出的它也为RNA结合蛋白质分类器的最具有分类性能的特征，并且它也发生在RNA结合蛋白质的结合位点处，因此可以推测“151”这个三联体是RNA结合蛋白质所特异的。其次为三联体特征“236”，对应为亲水大氨基酸（I, L, F, P）接着较低极性大氨基酸（Y, M, T, S）再接亲水大氨基酸（D, E），其他的依次类同。

上面的分析不仅提供了哪些“三联体”是对DNA结合蛋白质或者RNA结合蛋白质的识别起关键重要的特征，也为对这些蛋白的某些区域（例如，含有某种三联体）进行单点或者多点变异（例如，从某一类“三联体”变异到另一类“三联体”）后识别其功能起到了启示性作用。

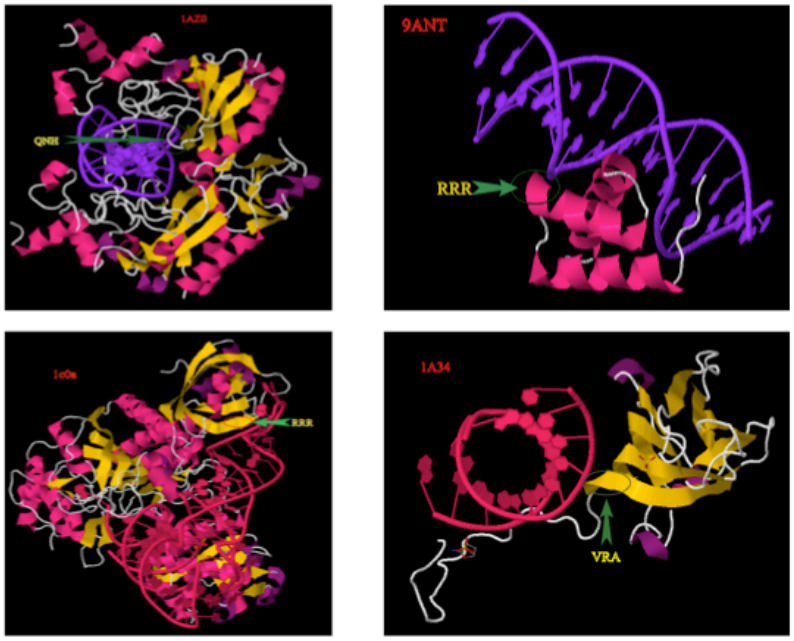


图2-6 部分DNA/RNA结合蛋白质分类器中重要的“三联体”特征图示

2.3.4 基于蛋白质全序列的分类器与基于蛋白质结构域的分类器

本章建立的模型是基于Swiss-prot的蛋白质全序列的数据的。除了用蛋白质全序列，还可以根据PDB数据库中获得的蛋白质序列（对应为蛋白质结构域或者部分实验测定的子序列，本章简称为“结构域”序列）构建分类模型。这时面临的一个问题是，针对同一种“三联体”编码方式，可否直接利用基于Swiss-prot的蛋白质全序列的数据构建的分类模型来预测PDB数据中对应的蛋白质子序列的数据呢？或者反过来，可否用基于PDB数据库中对应的蛋白质子序列数据构建的分类模型去预测全序列对应的蛋白质数据呢？

为此，首先专门收集了来自PDB数据库中的非冗余的DNA结合蛋白质数据集和相应的非结合蛋白质数据集，并基于“三联体”特征编码方式和支持向量分类机，构建了基于“结构域”的DNA结合蛋白质分类器。如在前节数据集中提到的，基于PDB数据库中的非冗余DNA结合蛋白质数据和非冗余非结合蛋白质数目分别为146个和250个。利用基于“留一法”的交叉验证方法，得到了基于蛋白质“结构域”或者子序列的DNA结合蛋白质分类器的分类性能，预测精度达到79%，相应的AUC取值高达0.86。这表明，基于蛋白质“结构域”或者子序列信息的“三联体”编码构建的分类模型能有较好的分类性能。

接着，利用两种类型（全序列和“结构域”或者子序列）数据构建的分类模型对彼此进行预测测试。由DNA结合蛋白质的全序列信息构建的分类器，预测仅由DNA结合蛋白质的“结构域”信息构建的测试集，得到的分类预测精度为61.4%，相应AUC取值为0.62。而若用由DNA结合蛋白质的“结构域”信息构建的分类器，对由DNA结合蛋白质的全序列信息构建的测试集进行预测，得到的分类预测精度仅为52.2%，相应AUC取值仅为0.54，具体结果参见图2-7。这个结果表明了想从基于全序列的分类模型去预测基于“结构域”序列的蛋白质数据，或者想从基于蛋白质“结

构域”序列的分类模型去预测基于全序列的蛋白质数据，都是不合适的。为了究其原因，考察了这两种类型（基于全序列和基于蛋白质“结构域”序列）的蛋白质数据集中，正、负类“样本点”（蛋白质）对应的“三联体”的分布，如图2-8。从图中可以看出，基于蛋白质“结构域”序列的“三联体”特征分布在“正、负”类间的分布差异较大，而在基于蛋白全序列的“正、负”类蛋白样本间的“三联体”特征分布则较为近似。这其中很大一部分原因可能是因为基于Swiss-prot数据库的蛋白质数据是蛋白质全序列数据，其有可能还包含其他类型的结构域。这也从侧面说明了利用仅基于DNA结合蛋白质“结构域”信息构建的分类器对应的分类性能，要比利用基于蛋白质全序列的数据构建的分类器性能要好。同时，也解释了为何基于这两种类型的数据集构建的分类器不能对彼此进行有效的预测。以上的结果分析表明，针对DNA结合蛋白质的识别问题，基于不同的数据类型（来自Swiss-prot的蛋白质全序列和来自PDB数据库中的蛋白质“结构域”或者蛋白质子序列），需要构建不同的分类器，两者之间不能有效地互换。

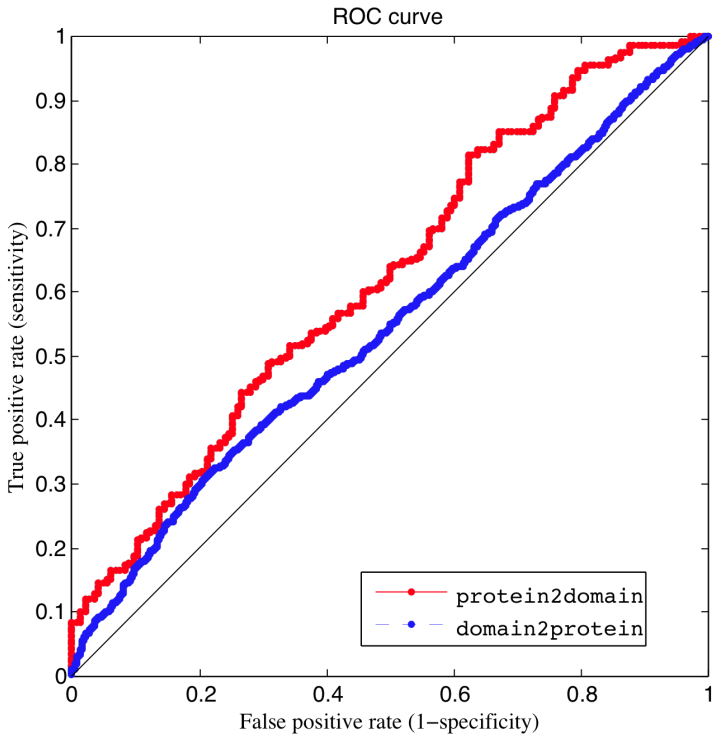


图2-7 由DNA结合蛋白质的不同序列编码的特征向量构建的分类器的分类性能比较

红色ROC曲线对应为由DNA结合蛋白质的全序列进行编码构建分类器，并对由DNA结合蛋白质的结构域序列编码得到的样本集进行预测的性能；蓝色ROC曲线对应为由DNA结合蛋白质的结构域序列进行编码构建分类器，并对由DNA结合蛋白质全序列编码得到的样本集进行预测的性能。

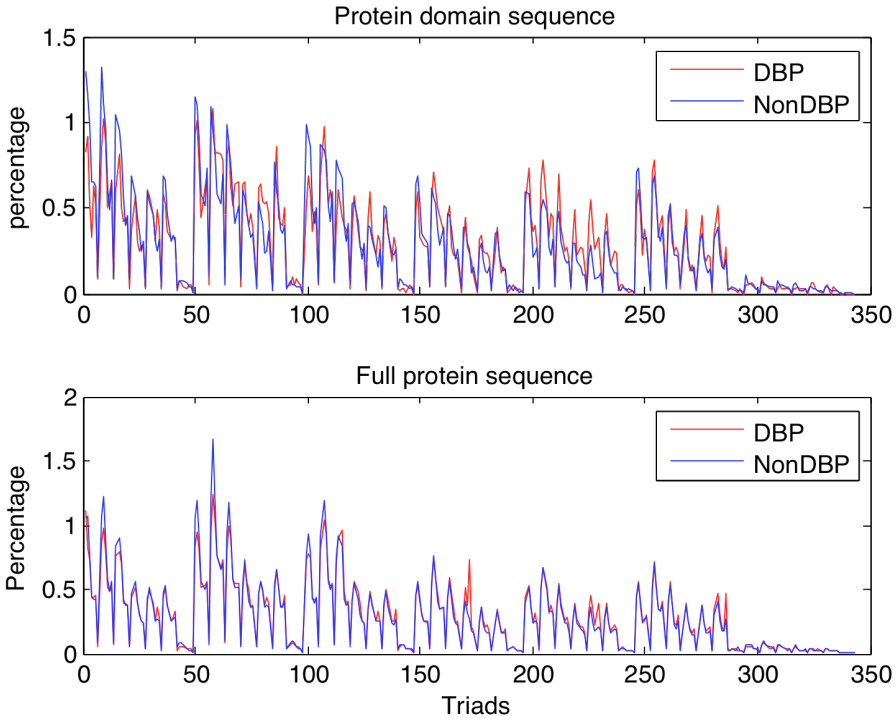


图2-8 两种不同蛋白质序列部分对应的三联体特征编码的分布

上图对应为基于蛋白质“结构域”或者子序列的三联体特征分布。下图对应为基于蛋白质全序列的三联体特征分布。基于蛋白质“结构域”或者子序列的“正、负类”样本间三联体特征分布差异较大。

2.4 结论和工作展望

本章尝试从序列的角度用新的特征属性——“三联体”特征编码，结合支持向量分类机实现了对DNA结合蛋白质和RNA结合蛋白质的预测。与前人的结果相比，新的特征编码对DNA结合蛋白质的预测较为有效，精度达到了78.93%，并且此种编码方式易于实现，为从序列角度实现对DNA结合蛋白质的预测提供了新的方法。同时本章所提出的“三联体”特征编码方式也显示了其能较好地区分DNA结合蛋白质和RNA结合蛋白质。通过统计比较不同的“三联体”对识别DNA结合蛋白质或者RNA结合蛋白质的有效性，发现很多具有显著特征的“三联体”大多位于DNA结合蛋白质或者RNA结合蛋白质与DNA或者RNA结合的结合表面。这表明“三联体”特征有效地表征了DNA结合蛋白质或者RNA结合蛋白质在结合位点处的结合模式。

通过分析本章已经取得的结果和不足，针对蛋白质与核酸之间的相互作用，未来值得进一步考虑和研究的内容仍有很多。

首先，对RNA结合蛋白质的识别需要进一步的改进。本章提出的“三联体”特征能较好地用来描述DNA结合蛋白质，并可以结合支持向量分类机较有效的识别DNA结合蛋白质。但是对于RNA结合蛋白质的识别仍旧没有展示出明显的优势。随着生物技术的发展和研究的深入，越来越多的文献表明生物体内的很多生物过程都可能和RNA与蛋白质的相互作用有关，从而对RNA结合蛋白质的识别也变得更为重要。与DNA的简单的双螺旋结构不同，RNA本身具有更为复杂的二级结构，所以对RNA结合蛋白质的识别也显得更为困难和复杂。

此外，从机器学习的角度而言，未来还可以考虑采用“集成学习”的策略[119]，将各个不同的分类器（诸如支持向量机、条件随机场、决策树等）加以整合、集成，减低因在特定训练集上得到的单个分类器可能引起的模型变差，从而使得最终的分类器能具有更强的稳健性和具有更强的预测性能。

考虑预测DNA结合蛋白质或者RNA结合蛋白质的同时，另一个值得考虑的问题是DNA结合蛋白质或者RNA结合蛋白质的结合位点[120]。在这方面，目前也已经有很多学者进行了研究[121-125]，但是当前发布的对位点预测的工具之间往往存在着对同一结合蛋白质预测出较为不同的结合位点的结果，表明对位点预测问题仍需要进一步的分析和研究。此外，研究发现在某些DNA结合位点处发生单点变异即有可能导致疾病的发生（例如，P53蛋白的某些位点处发现变异[126]）。知道DNA结合蛋白质或者RNA结合蛋白质的具体的结合位点，可以加深人们对生物过程的理解，因为一般认为只有这些位点才是真正起生物学功能的位点。同时还可以指导人们对DNA结合蛋白质或者RNA结合蛋白质进行“点变异”进而影响其生物功能（例如，可以实现新转录因子的人工合成，为生物制药提供必要的辅助等）[127]。从序列角度识别哪些是DNA或者RNA结合蛋白质，并且同时进一步识别它们的结合位点将是未来设计整合预测分类器的一个重要方向。

本章主要实现了从蛋白质序列的角度来预测DNA结合蛋白质和RNA结合蛋白质。可以说，针对蛋白质与核酸相互作用研究而言，仅关注了问题的一个方面，还需要进一步的观察给定DNA或者RNA结合蛋白质能与DNA或者RNA的哪一段进行相互作用。加拿大多伦多大学学者Tim Hughes等人测得了小鼠中（部分）168个含homeo-domains的DNA结合蛋白质与DNA片段相互作用的信息（转化为Z-score）[128, 129]，为进一步研究DNA结合蛋白质与DNA序列片段之间的相互作用预测提供了新的宝贵的数据资源。同时，利用现有的针对DNA结合蛋白质与DNA片段之间相互作用亲和力的数据，设计有效的预测模型，实现对未知DNA结合蛋白质与DNA片段之间相互作用强弱的预测，将是未来的一个工作重点。此外，针对RNA结合蛋白质，也有实验测得了部分RNA结合蛋白质的结合模体，确知它们与哪些RNA的片段结合[130]，有效利用这部分数据建立预测模型推断RNA结合蛋白质的结合特异性也将是本章研究内容的一个重要延伸。

本章工作仅仅只是为确定整个蛋白质与核酸分子相互作用网络（例如转录调控网络（Transcriptional networks））的一个基本铺垫。随着大规模ChIP-seq数据的获取，将为未来研究蛋白质-DNA和蛋白质-RNA之间的相互作用提供更多的数据支持，为研究蛋白质与核酸分子相互作用网络提供保障。

第三章 WW结构域-短肽相互作用预测研究

本章开展对蛋白质与蛋白质相互作用预测的研究，并且主要从蛋白质结构域与短肽相互作用的角度进行研究。很多蛋白质是由单个或者多个结构域组成的，它们之间的相互作用大部分都是由其中的结构域与结构域或者结构域与短肽配体相互作用而完成的。在结构域层面上研究蛋白质结构域与蛋白质短肽配体之间的相互作用，不仅可以更为深入地了解蛋白质与蛋白质相互作用的分子机制，而且可以更为全面地发现潜在的蛋白质结构域-短肽相互作用对，从而可以通过针对各种不同的结构域构建的结构域及其潜在的短肽相互作用网络，更精确地实现整个蛋白质组层面的相互作用网络图谱的绘制。目前，虽然已经有研究工作关注在结构域层面绘制结构域-短肽之间的相互作用网络，但是通过高通量实验手段实现蛋白质结构域和蛋白质短肽之间相互作用测定仍存在很大困难，同时相关的计算预测模型也相对缺乏。本章针对信号传导中最为常见的结构域之一的WW结构域进行研究，从计算模型角度入手，开展构建预测WW结构域与短肽之间是否发生相互作用以及相互作用强弱水平方面的模型的研究工作。

3.1 引言

WW结构域是最为常见的信号传导中的结构域之一。WW结构域的序列长度较小，大概由40个左右的氨基酸构成。该序列通常由3个β片段折叠而成。该结构域由于在距离长度约20-23个氨基酸的两个位置上大多含有色氨酸（W）而得名，是美国学者Marius Sudol博士于1994年发现的[131]。WW结构域的图示见图3.1。该结构域可以多个（≤4）出现在某些蛋白质中，其中比较典型的几个含有WW结构域的蛋白质结构域组成图可参见图3.2。WW结构域分布广泛，它既存在于单细胞生物体内，又见于多细胞生物体中。据SMART数据库（SMART6）[7, 132]统计显示，目前在生物界内存在约3625个WW结构域，分别存在于2153个蛋白质中，其中人类大约含有260个WW结构域（据Pfam统计），对应在64个人类蛋白质中。含有WW结构域的蛋白质生物功能多样化，具有结构功能、调控功能、信号传导功能等。而WW结构域的生物功能主要是通过与其他蛋白质的相互作用来实现，即主要是通过识别富含有脯氨酸（P）的配体蛋白质来实现其功能。当前，WW结构域的配体识别模式主要是以下4种[133]：PPXY模体、PPLP模体、PRP模体以及识别磷酸化的（S/T）P模体，其中X为任意氨基酸。大量的实验表明，通常WW结构域结合短肽配体的相互作用强度（电离常数K_d值）在*u*M范围内。越来越多的研究发现WW结构域所介导的信号传导系统常常与很多疾病有关，比如利德尔（Liddle）高血压综合症、杜氏-贝克尔（Duchenn-Becker）肌肉萎缩、阿尔茨海默氏症（Alzheimer's）、亨廷顿（Huntington）疾病、甚至癌症等[134]。正是由于WW结构域与诸多疾病有关，具有如此重要的研究价值，越来越多的学者开始致力于开展与WW结构域相关的研究。其中在蛋白质组层面上研究WW结构域与其结合的配体之间的相互作用是当前研究WW结构域功能的一个重要方向，也是蛋白质相互作用研究领域的一个热点。

针对WW结构域与其短肽相互作用的研究，已有不少学者分别从实验和计算的角度开展了工作。当前测定WW结构域与其结合配体之间的相互作用的方法和相关研究工作主要有如下几个方

面的内容。

在实验方法方面，主要有1）短肽微阵列高通量实验手段。Hu等人于2004年，通过整合短肽的合成、蛋白质微阵列以及利用生物信息学的手段进行高通量的筛选技术完成了对人体中65个WW结构域的相互作用图谱的绘制[88]。由于该实验是基于现有结合模体筛选的短肽序列，所以此方法不易发现WW结构域的新的识别模体。2）蛋白质芯片技术。Hesselberth等人在2006年发表了利用蛋白质芯片技术测定WW结构域与其他蛋白质之间的相互作用谱图[90]。该实验得到的蛋白质相互作用网络图谱是对WW结构域相互作用网络的一个有利的补充。但是由于该实验是基于合成蛋白质的信息测得的相互作用，并不能精确地提供是相应蛋白质中的哪一段短肽与特定的WW结构域相互作用。3）蛋白质质谱技术。Ingham等人于2005年通过蛋白质质谱技术测定了WW结构域-配体混合体[89]。该技术并不依赖前人得到的蛋白质短肽配体信息，为测得新的WW结构域-短肽配体相互作用提供了大量的互补信息。需要注意的是，该蛋白质质谱技术得到的WW结构域与配体相互作用复合体含有了并未直接与WW结构域发生物理相互作用的配体。4）NMR谱方法。Otte等人于2003年利用NMR技术结合短肽scanning技术测得了42个WW结构域的短肽相互作用特性[87]。值得注意的是，该文的一个工作表明了有些WW结构域-短肽相互作用在短肽配体发生磷酸化时才真正起作用（主要以人类的PIN1和酵母的ESS1中的WW结构域为例），而在未被磷酸化之前，该短肽配体则不与相应的WW结构域进行相互作用。有趣的是，只是含有丝氨酸或者色氨酸的短肽被磷酸化时才可能发生相互作用，对于含有酪氨酸的短肽即使发生了磷酸化也未有相互作用。

虽然，通过实验的方法可以得到相互作用的数据，但是往往实验的方法是既费时又耗财的。近年来，随着生物数据的积累和生物信息技术的发展，使得通过计算的方法来预测结构域和配体之间的相互作用成为一个主要工具。

目前基于计算模型进行预测WW结构域与短肽配体相互作用的工作仍还不多。其中比较典型的是Hu等人的文章中提到了利用特异性打分矩阵（PWM或者PSSM）方法对其中49个WW结构域进行筛选潜在的短肽配体[88]。此外，2004年，Wade小组通过结构模拟，同源对比的方法实现了对WW结构域与其配体相互作用的预测[135]。但是这种基于结构的计算方法需要有结构信息，计算时间较长。而且该工作并没有真正意义上的基于高通量的数据来获取预测模型。

上述实验获取的WW结构域与短肽相互作用数据中，有基于蛋白质层面的相互作用的二值数据（是否发生相互作用），也有专门针对WW结构域的相互作用强弱的数据。随着越来越多的相互作用数据的测得，人们开始越来越关注相互作用蛋白质对之间的亲和力的强弱。其中Hu等人在2004年测得的关于人类蛋白质中65个WW结构域与其配体相互作用图谱的工作中，就获得了第二种类型的数据。在该工作中，不但确定了该65个WW结构域与哪些配体相互作用，而且还测得了某种程度上表征它们之间相互作用的强弱的“亲和力”（即该实验得到的AU值）。

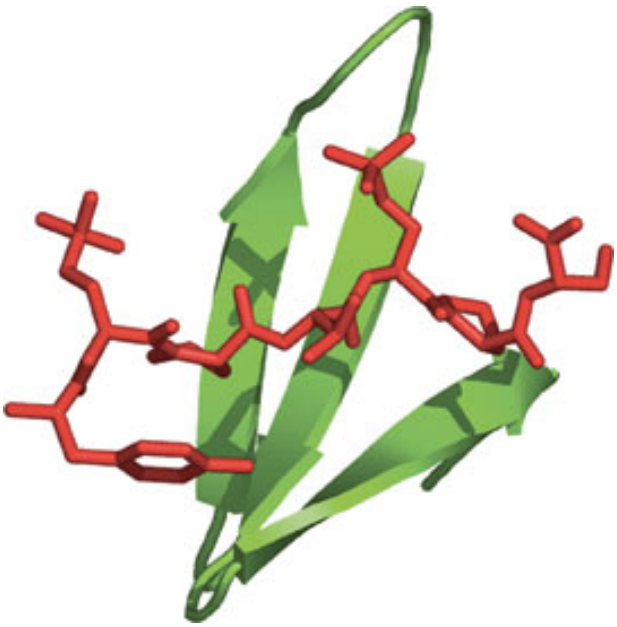


图3-1 WW结构示意图

绿色代表WW结构域，红色代表配体（Ligand），图摘自http://www.cellsignal.com/reference/domain/ww.html

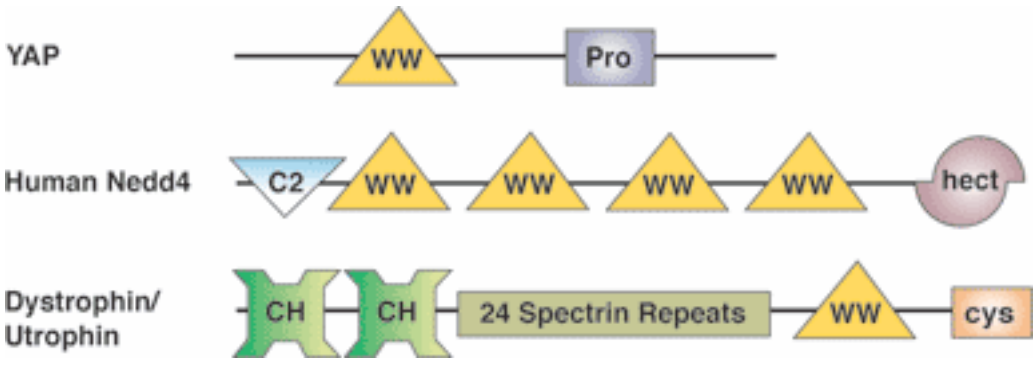


图3-2 含有WW结构域的蛋白质结构域组成图示

图摘自http://www.cellsignal.com/reference/domain/ww.html

综上所述，目前从计算模型的角度开展针对WW结构域-短肽相互作用预测的工作已显得非常急迫而重要。与此同时，高通量数据的产生为我们设计高效的基于机器学习的预测模型提供了非常重要的资源和保障。本章所研究的数据集包括上述65个WW结构域的数据，加之后来更新的13个WW结构域，总共为78个人类WW结构域的短肽相互作用数据集。基于上述数据，本章建立了可以预测WW结构域与配体之间是否发生相互作用以及相互作用强弱水平的计算模型。利用该模型，可以对人类蛋白质组进行大规模的筛选，以发现潜在的可能与WW结构域相互作用的配体蛋白，并给出其强弱。同时基于此模型，可以对其他物种中的WW结构域预测与其相互作用的配体蛋白质及其强弱水平。

3.2 数据集和方法

3.2.1 数据集

本章用到的WW结构域-短肽相互作用的数据来自于文章[88]以及后来更新的数据。Hu等人测得该数据所采用的实验手段的具体表述如下；

步骤一，含有WW结构域的GST融合蛋白质合成；通过PCR对WW结构域对应的cDNA序列进行增扩。WW结构域通过NH_2端进行结合GST，同时在*E. Coli*中进行表达。

步骤二，短肽配体序列合成；所有的短肽均利用Mimotopes SynPhase公司的技术产品进行合成，并置放于具有96孔的微阵列盘中。之后在MeOH或者DMF中洗涤，接着利用高通量技术LC-MS进行分析处理，最后再通过Gilson(Middleton, WI, USA)公司的HPLC系统进行纯化。

步骤三，构建交叉亲和矩阵；在每张96孔盘上均有合成的短肽序列，每一张盘对应一个WW结构域，通过microplate absorbance阅读器测得该WW结构域与该盘上的短肽之间的相互作用亲和力，用absorbance unit（AU）值来表示相互作用强弱。

具体实验过程如图3-3所示。

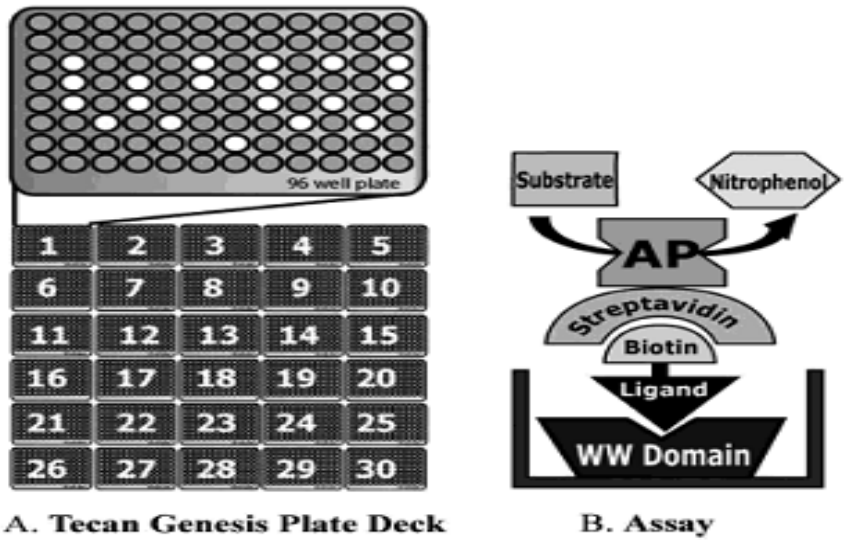


图3-3 WW结构域-短肽相互作用测定实验流程

（A）TECAN微阵列盘图示。每一张盘对应一个WW结构域，每96个孔的微阵列盘装有96个短肽。（B）实验装置。阅读器测定每一个WW结构域探针识别的带有生物素的短肽序列的相互作用亲和力（AU值）。该图摘自[88]。

实验中合成的短肽序列是人类蛋白质中的真实序列，是按先前文献报道的几类模体（例如，较为保守的‘PPXY’模体）进行筛选后得到的。从Swiss-prot和TrEMBL数据库中查询含有上述几类保守模体的序列，一旦匹配就选为潜在的短肽。同时要求合成的短肽配体序列尽可能地包括保守模体上游N端的4个氨基酸和下游C端的4个氨基酸。由于某些模体具有多于4个氨基酸长度的保守片段，且有时有些短肽序列刚好位于蛋白序列的N端或者C端，不能保证具有4个长度的上下游（Flanking region）序列，所以最后筛选得到的短肽序列的长度介于10-16个氨基酸长度之间。在Hu等人的文中，从Swiss-prot和TrEMBL数据库中找到了匹配的2189个短肽，实际合成了1930个。在Hu等人的工作中，一共65个人类WW结构域被识别、合成和分析。在后续的工作中，我们重新合成了13个人类WW结构域，同时重新合成了611个短肽。为了方便计算模型的建立，这里仅仅考虑长度为12个氨基酸的短肽序列，这部分短肽占了全部短肽序列的96%，所以下面建立的模型仍具有较强的普适性。在最后更新的数据中涵盖了78个WW结构域和2428个人类短肽配体序列，总共测得了63783对WW结构域-短肽相互作用对。由于实验条件的因素，在78×2428规模的交叉亲和矩阵中，仍有大部分的WW结构域-短肽相互作用对（约占67%）之间的AU值未测得，具体的分布图见图3-4。

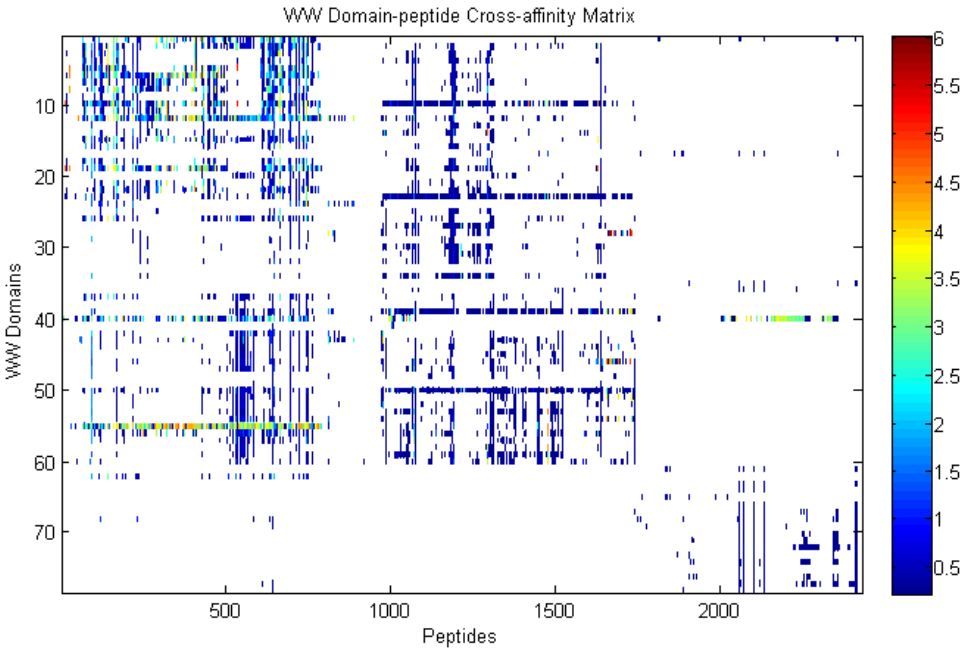


图3-4 WW结构域-短肽相互作用“亲和力”交叉矩阵图示

图中颜色代表相应WW结构域-短肽对的亲和力大小，由蓝色至红色代表亲和力AU值由小到大的变化，而空白处代表该对WW结构域-短肽对未测得亲和力AU值。

在上述得到的63783个WW结构域-短肽对中，亲和力（AU）的取值从0.21到6不等。在对该数据集进行数据分析、建模之前，需要合理地确定正类相互作用对和负类相互作用对的阈值。如文[88]中Hu等人所分析的，亲和力AU值的分布呈现了3个较为明显的区域，尤其在AU=0.5处有一个很大的转变，所以取AU=0.5是一个比较合理的阈值（如图3-5所示，图中呈现了3个不同的AU值区域，在AU=0.5处有个突变）。本章也采用AU=0.5作为正类、负类样本区别的一个阈值。事实

上，对这63783个WW结构域-短肽对的AU值进行统计发现，64.9%的WW结构域-短肽对的AU值不高于0.5；19.8%的介于0.5～1.5之间；7.1%的介于1.5～2.5之间；4.4%的介于2.5～3.5之间；仅有3.8%的WW结构域-短肽对的AU值高于3.5；具体分布如图3-6。根据此分布，可以粗略地将WW结构域与短肽蛋白相互作用对按其亲和力AU值的大小划分为以下5大类：负类定义为亲和力<0.5AU；较弱的相互作用对定义为亲和力介于0.5～1.5AU之间；弱相互作用对亲和力介于1.5～2.5AU之间；一般相互作用对亲和力介于2.5～3.5AU之间；强相互作用对定义为亲和力大于3.5AU。

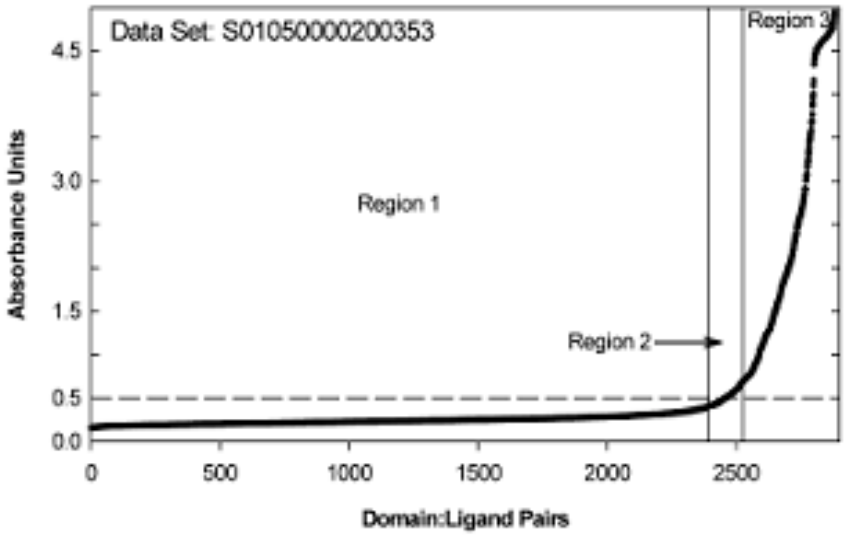


图3-5 WW结构域-短肽相互作用亲和力之AU值排序图示

（该图摘自文[88]）

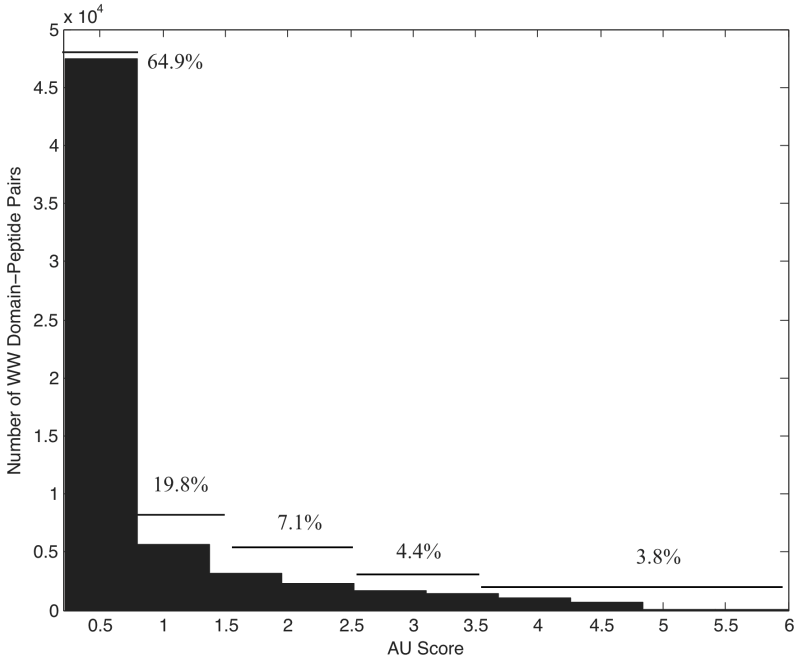


图3-6 WW结构域-短肽相互作用对之AU值分布

3.2.2 WW结构域蛋白质列表

WW domain Name	ACCESS NUMBER	Short name	Long name
D00001	Q9H0M0	WWP1_1	NEDD4-like E3 ubiquitin-protein ligase WWP1, WW1
D00002	Q9H0M0	WWP1_2	NEDD4-like E3 ubiquitin-protein ligase WWP1, WW2
D00003	Q9H0M0	WWP1_3	NEDD4-like E3 ubiquitin-protein ligase WWP1, WW3
D00004	Q9H0M0	WWP1_4	NEDD4-like E3 ubiquitin-protein ligase WWP1, WW4
D00005	O00308	WWP2_1	NEDD4-like E3 ubiquitin-protein ligase WWP2, WW1
D00007	O00308	WWP2_3	NEDD4-like E3 ubiquitin-protein ligase WWP2, WW3
D00008	O00308	WWP2_4	NEDD4-like E3 ubiquitin-protein ligase WWP2, WW4
D00009	Q96QZ7	MAGI1_1	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1, WW1
D00016	Q96PU5	NED4L_2	E3 ubiquitin-protein ligase NEDD4-like, WW2
D00017	Q96PU5	NED4L_3	E3 ubiquitin-protein ligase NEDD4-like, WW3
D00018	Q96PU5	NED4L_4	E3 ubiquitin-protein ligase NEDD4-like, WW4
D00029	Q96QZ7	MAGI1_2	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 1, WW2
D00073	Q86UL8	MAGI2_1	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 2, WW1
D00074	Q86UL8	MAGI2_2	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 2, WW2
D00075	Q96J02	AIP4_1	E3 ubiquitin-protein ligase Itchy homolog, WW1
D00076	Q96J02	AIP4_2	E3 ubiquitin-protein ligase Itchy homolog, WW2
D00077	Q96J02	AIP4_3	E3 ubiquitin-protein ligase Itchy homolog, WW3
D00078	Q96J02	AIP4_4	E3 ubiquitin-protein ligase Itchy homolog, WW4
D00084	Q5TCQ9	MAGI3_1	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 3, WW1
D00086	P46934	NEDD4_1	E3 ubiquitin-protein ligase NEDD4, WW1
D00087	P46934	NEDD4_2	E3 ubiquitin-protein ligase NEDD4, WW2
D00088	P46934	NEDD4_3	E3 ubiquitin-protein ligase NEDD4, WW3
D00089	P46934	NEDD4_4	E3 ubiquitin-protein ligase NEDD4, WW4
D00139	Q5TCQ9	MAGI3_2	Membrane-associated guanylate kinase, WW and PDZ domain-containing protein 3, WW2
D00185	P46937	YAP1_1	65 kDa Yes-associated protein, WW1

D00186	Q76N89	HECW1_1	E3 ubiquitin-protein ligase HECW1, WW1
D00187	Q76N89	HECW1_2	E3 ubiquitin-protein ligase HECW1, WW2
D00188	Q9BYW2	SETD2	Histone-lysine N-methyltransferase SETD2
D00189	O00213	APBB1	Amyloid beta A4 precursor protein-binding family B member 1
D00190	Q8N1G2	FTSJD2	S-adenosyl-L-methionine-dependent methyltransferase FTSJD2
D00191	O75400	PRPF40A_1	Pre-mRNA-processing factor 40 homolog A, WW1
D00192	O75400	PRPF40A_2	Pre-mRNA-processing factor 40 homolog A, WW2
D00193	P46940	IQGAP1	Ras GTPase-activating-like protein IQGAP1
D00194	P11532	Dystrophin	Dystrophin
D00195	P46939	Utrophin	Utrophin
D00196	O60861	GAS7	Growth arrest-specific protein 7
D00197	Q96PU5	NEDD4L_1	E3 ubiquitin-protein ligase NEDD4-like, WW1
D00215	O14776	TCERG1_1	Transcription elongation regulator 1, WW1
D00216	O14776	TCERG1_2	Transcription elongation regulator 1, WW2
D00230	Q8N3X1	FNBP4_1	Formin-binding protein 4 , WW1
D00231	O60828	PQBP1	Polyglutamine-binding protein 1
D00232	O75554	WBP4_1	WW domain-binding protein 4, WW1
D00233	O75554	WBP4_2	WW domain-binding protein 4, WW2
D00241	Q6NWY9	PRPF40B_1	Pre-mRNA-processing factor 40 homolog B, WW1
D00242	Q6NWY9	PRPF40B_2	Pre-mRNA-processing factor 40 homolog B, WW2
D00243	Q9BTA9	WAC	WW domain-containing adapter protein with coiled-coil
D00244	Q8IWW6	ARHGAP12_2	Rho GTPase-activating protein 12, WW2
D00245	Q6ZUM4	ARHGAP27_3	Rho GTPase-activating protein 27, WW3
D00251	Q9HCE7	SMURF1_1	E3 ubiquitin-protein ligase SMURF1, WW1
D00252	Q9HCE7	SMURF1_2	E3 ubiquitin-protein ligase SMURF1, WW2
D00253	Q92870	APBB2	Amyloid beta A4 precursor protein-binding family B member 2
D00254	Q13576	IQGAP2	Ras GTPase-activating-like protein IQGAP2
D00255	Q13474	DRP2	Dystrophin-related protein 2

D00256	O95704	APBB3	Amyloid beta A4 precursor protein-binding family B member 3
D00348	O95817	BAG3_1	BAG family molecular chaperone regulator 3, WW1
D00349	Q9NZC7	WWOX_1	WW domain-containing oxidoreductase, WW1
D00350	Q9NZC7	WWOX_2	WW domain-containing oxidoreductase, WW2
D00351	Q8N3X1	FNBP4_2	Formin-binding protein 4, WW2
D00352	O15428	PIN1L	Putative PIN1-like protein
D00353	Q9UPV0	CEP164	Centrosomal protein of 164 kDa
D00354	Q13526	PIN1	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1
D00357	Q86U42	PABPN1	Polyadenylate-binding protein 2
D00561	P09619	PDGFRB	Beta-type platelet-derived growth factor receptor
D00563	P56539	CAV3	Caveolin-3
D00699	Q9C0H5	ARHGAP39_2	Uncharacterized protein KIAA1688, WW2
D00700	Q9P2P5	HECW2_2	E3 ubiquitin-protein ligase HECW2, WW2
D00701	Q9BTA9	WAC_a	WW domain-containing adapter protein with coiled-coil
D00702	Q9GZV5	WWTR1	WW domain-containing transcription regulator protein 1
D00703	Q9HAU4	SMURF2_3	E3 ubiquitin-protein ligase SMURF2, WW3
D00704	Q9H4B6	SAV1_2	Protein salvador homolog 1, WW2
D00705	Q9BRR9	ARHGAP9	Rho GTPase-activating protein 9
D00706	Q8WYQ5	DGCR8	Microprocessor complex subunit DGCR8
D00708	Q9HAU4	SMURF2_1	E3 ubiquitin-protein ligase SMURF2, WW1
D00710	Q9H4Z3	PCIF1	Phosphorylated CTD-interacting factor 1
D00711	Q9H4B6	SAV1_1	Protein salvador homolog 1, WW1
D00712	Q9HAU0	PLEKHA5_2	Pleckstrin homology domain-containing family A member 5, WW2
D00713	Q8IX03	WWC1_1	Protein WWC1, WW1
D00714	Q8IX03	WWC1_2	Protein WWC1, WW2

3.2.3 特征编码

在用支持向量机等机器学习模型之前，需要对WW结构域-短肽对进行特征编码。通常需要用一个数值型的向量来描述。由于对WW结构域-短肽对进行编码，涉及到WW结构域和短肽两个部分，可以对各自两个部分分别进行编码，然后组合起来，也可以直接从WW结构域-短肽对入手从整体上设计编码方式。

针对第一类编码方式，本章采用如下两种编码方式：

1. 二值正交编码：即对WW结构域序列或者短肽序列的每个氨基酸采用20维的二值向量来描述。首先将20种氨基酸按某种顺序排好序，然后在20维的向量中相应的位置赋值1，其他位置为0。例如氨基酸A可以描述为(1, 0, 0, …, 0)。这样就可以分别对WW结构域序列和短肽序列将序列中相应的氨基酸进行编码。在各自得到各自的二值编码之后，可以将两者按照WW结构域编码向量在前，短肽序列编码向量在后面的顺序结合得到一个新的向量。对于WW序列，可以采用全序列，也可以采用部分功能位点（例如物理结合位点，binding sites）构成的伪序列来代表。

注：在此编码方式中要求各个WW结构域的序列长度是一致的，这就需要进行多序列比对。本章将采用Mafft软件[136]对78个WW结构域进行多序列比对，具体多序列比对的结果参见图3-7。

2. 对WW结构域利用Profeat[112]进行编码，该方法可以获取蛋白序列的氨基酸组成，二级结构信息组成，各个氨基酸的物理化学特性等属性，也即主要是采用Profeat工具中的group5的特征组。由于短肽序列太短，不适于采用Profeat进行编码，所以仍采用正交二值编码处理。

针对第二类编码方式，本章采用：

3. 基于WW结构域-短肽相互作用对中物理接触的“接触对”来实现数值编码。

首先，从现有的PDB数据库[110]下载到10个涉及到同时包含人类WW结构域以及短肽的蛋白复合物的结构文件，具体信息见表3-1。如果蛋白质结构是由NMR技术测得，对应会有多个模型，这里只取第一个作为最优模型来获取各个氨基酸的三维坐标。然后针对每一个蛋白质结构，计算WW结构域中的每一个氨基酸的主α碳原子与短肽序列中每一个氨基酸的主α碳原子直接的距离。当两个主α碳原子之间的距离小于给定阈值时（从3.5Å增加至6Å，步长为1Å），定义为该两个氨基酸是物理相互接触的。由此，提取出10个蛋白质结构中所有满足阈值的物理接触对。本章最终取定5Å为阈值，10个蛋白对应的相应物理接触对如图3-8所示。

其次，在获得物理接触对之后，就可以通过这些物理接触的氨基酸对来描述整个WW结构域-短肽相互作用对。按如下的标准选取物理接触对作为最终的氨基酸对集合：只要某接触对在其中两个蛋白质结构中出现过就被选中。以此来表示WW结构域-短肽相互作用对，最终得到了52个WW结构域-短肽相互作用对的物理接触对，涉及到多序列对比之后的WW结构域序列中的14个结合位点和短肽序列中的9个位置。

表3-1 含有WW结构域-短肽的蛋白质复合物之PDB数据信息

PDB ID	实验方法	WW domain	Peptide
1I5H	SOLUTION NMR	GSPVDSNDLGPLPPGWEERTHTDGRVFFINHNIKKTQWEDPRMQNVAITG	GSTLPIPGTPPPNYDSL
1JMQ	SOLUTION NMR	FEIPDDVPLPAGWEMAKTSSGQRYFKNHIDQTTTWQDPRKAMLSQM	GTPPPPYTVG
1YWI	SOLUTION NMR	GSRRASVGSAKSMWTEHKSPDGRTYYYNTETKQSTWEKPDD	APPTPPPLPP
2DJY	SOLUTION NMR	GPLGSGPLPPGWEIRNTATGRVYFVDHNNRTTQFTDPRLSAN	GPLGSELESPPPPYSRYPMD
2DYF	SOLUTION NMR	GSWTEHKSPDGRTYYYNTETKQSTWEKPDD	GSTAPPLPR
2EZ5	SOLUTION NMR	GPLGSGEEEPLPPRWSMQVAPNGRTFFIDHASRRTTWIDPRNGRAS	TGLPSYDEALH
2HO2	X-RAY DIFFRACTION	GSDLPAGWMRVQDTSGTYYWHIPTGTTQWEPPGRASPS	PPPPPPPPPL
2JMF	SOLUTION NMR	GPLGSPEFHMVSLINEGPLPPGWEIRYTAAGERFFVDHNTRRTTFEDPRPGAP	GPLGSPNTGAKQPPSYEDCIK
2JO9	SOLUTION NMR	GAMGPLPPGWEKRTDSNGRVYFVNHNTRITQWEDPRS	EEPPPPYED
2OEI	X-RAY DIFFRACTION	GSDLPAGWMRVQDTSGTYYWHIPTGTTQWEPPGRASPS	PPPPPPLPP

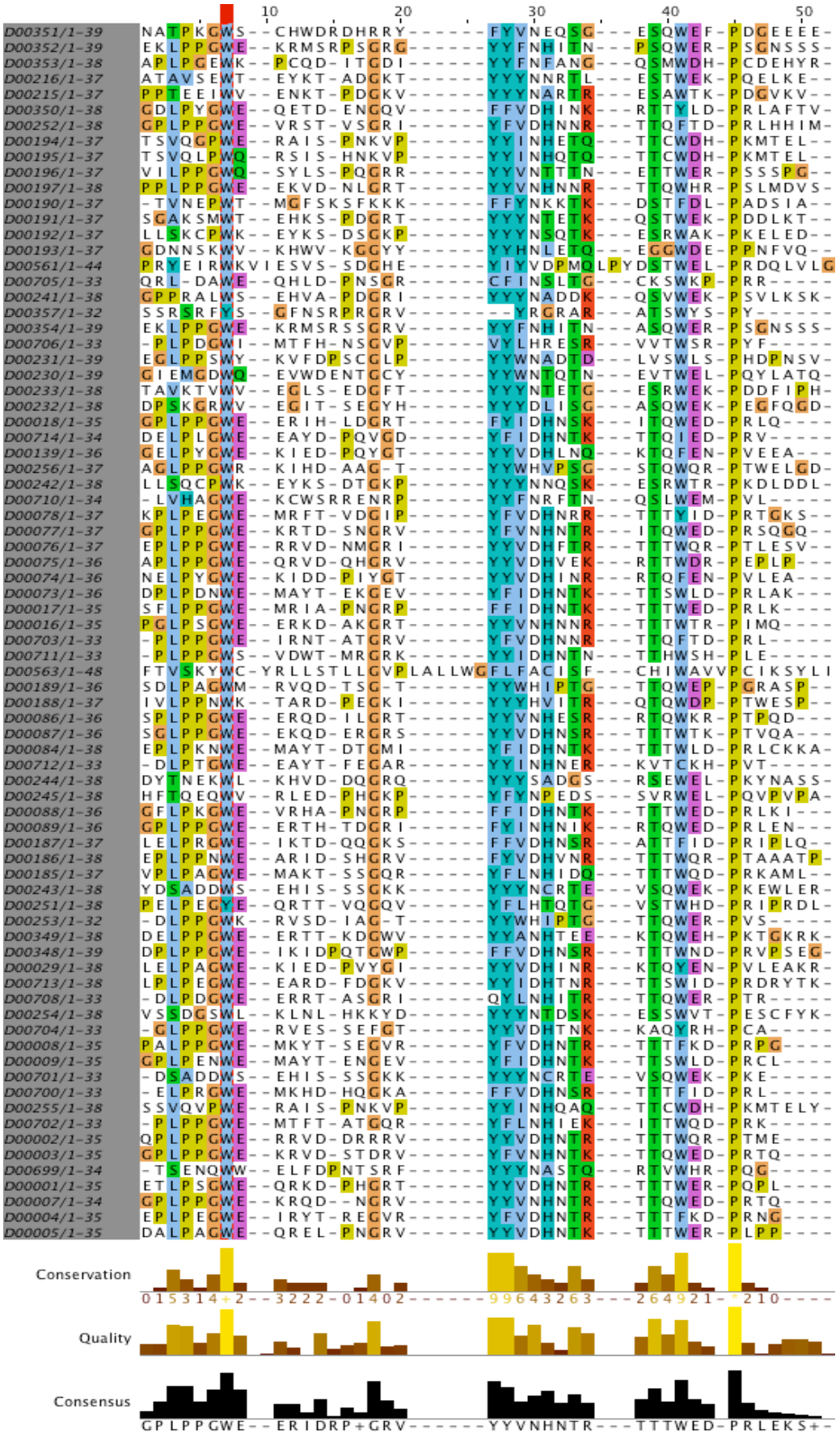


图3-7 WW结构域多序列对比结果

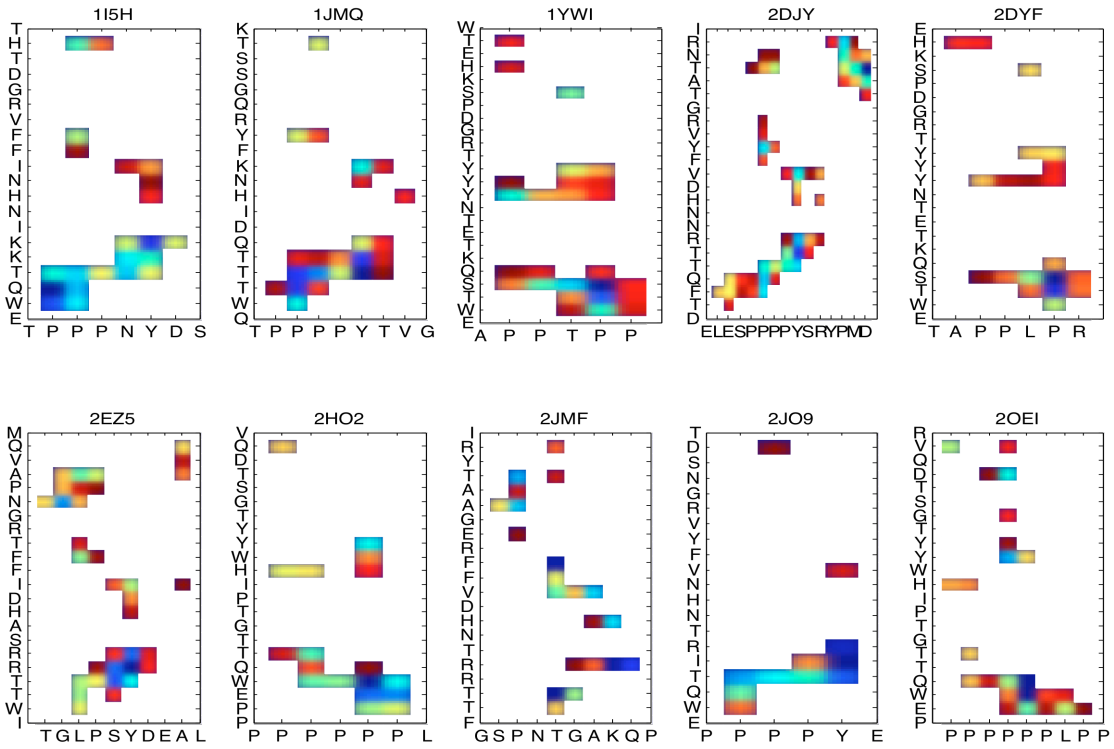


图3-8 WW结构域-短肽物理相互接触对

图中颜色由蓝色至红色代表相应氨基酸中主α碳原子之间的距离由小至大的值。白色部分代表WW结构域和短肽序列中对应的两个主α碳原子之间的距离小于5Å。

最后，采用如下两种方式来编码上述得到的所有的物理接触氨基酸对：正交二值编码方式和成对数值化“PAIR”方式。对于正交二值化编码方式，因为任何一个氨基酸对都含有两个氨基酸，所以对任意一对氨基酸对都需要用一个400维的向量来表示，即在对应的氨基酸对位置上取值为1，其他均为0。对于成对数值化“PAIR”方式，采用如文[73]中所提到的方式，具体定义数值向量过程如下：

1. 针对前面得到的52个氨基酸接触对，记任意一对氨基酸接触对为 $(d_i,p_j)_k$，i=1,...14, j=1,...,9, k=1,...,52。其中 d_i和 p_j 分别代表WW结构域中的第i个接触位点和短肽序列中第j个接触位点上对应的氨基酸，k则代表第k对氨基酸物理接触对。

2. 针对上述52个氨基酸接触对，定义任意一个氨基酸对（400种组合）在上述各个氨基酸接触对中，出现在“正、负”类集合中的偏好。为此，首先分别定义频数向量 $f_k^+(d,p)$和 $f_k^-(d,p)$：

$$f_k^+(d,p)=\frac{n_k^+(d,p)}{N_k^+}\text{和}\ f_k^-(d,p)=\frac{n_k^-(d,p)}{N_k^-},\tag{3-1}$$

其中 $n_k^+(d,p)$和 $n_k^-(d,p)$分别代表在正类和负类相互作用对中 $(d,p)_k$ 对出现的次数，k仍代表第k对氨基酸物理接触对。然后再将各个氨基酸组合对根据在“正、负”类中出现的上述频数进行数

值化，以反映其对WW结构域-短肽对是否发生相互作用的特异性，具体如下：

$$(d,p)_k \rightarrow C_k(d,p) = \frac{f_k^+(d,p) - f_k^-(d,p)}{\max(f_k^+, f_k^-)}。 \quad (3\text{-}2)$$

显然，C值介于-1到+1之间。具体的数值化结果示意图可见图3-9。

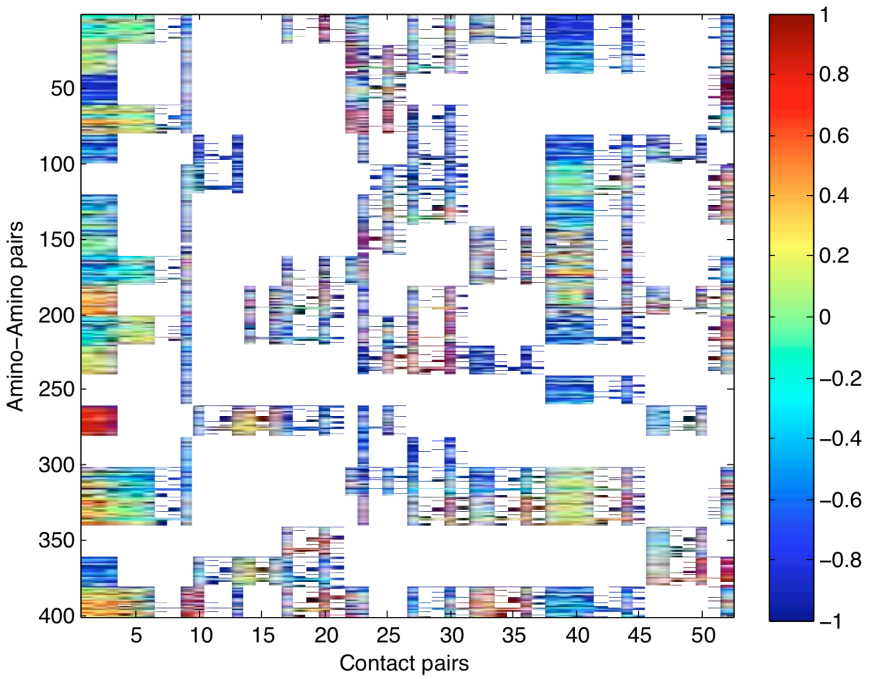


图3-9 氨基酸物理接触对之数值化编码

图中横坐标代表52个氨基酸物理接触对，纵坐标代表400种可能的氨基酸对。图中每一个元素代表某一对氨基酸对在给定物理接触对上的数值化取值（代表结合特异性偏好），其中白色部分代表相应的氨基酸对在该物理接触对上未曾出现。

3.2.4 支持向量分类机和回归机

本章采用支持向量分类机和支持向量回归机来进行建模。我们仍采用LibSVM作为实现支持向量回归机模型的软件，其中涉及到的参数利用网格搜索方法来搜寻最优参数。

在此节关注两分类问题，即要回答给定一个WW结构域-短肽蛋白对，预测它们之间是否发生相互作用。有两种策略可以处理这个问题。第一种，鉴于现有的训练集中的任何一个WW结构域-短肽对都对应测得一个AU值，可以利用这些AU值的信息结合支持向量回归机建立一个回归模型，然后根据选定的阈值如0.5AU来判定正类或者负类。另外一种策略是，先将给定的训练集对应的AU值离散化到正类、负类（对应>0.5AU和≤0.5AU），然后再结合支持向量分类机建立分类模型，通过该模型来判定测试的WW结构域-短肽对的正负类类别归属。

3.2.5 不均衡样本处理策略

在训练分类模型的时候，会遇到正类点和负类点样本不均衡的情况（如针对本章的数据集，如果采用不同的AU值作为定义正类点和负类点的阈值，就会出现负类点比正类点多很多的情形，

下文将对此进行具体讨论）。通常，由于样本不均衡会导致最终学到的模型具有偏向性，所以需要采用合适的策略来避免样本不均衡对模型训练所带来的影响。本章将采用一种基于“欠抽样”的抽样策略的变型来处理样本不均衡问题。具体地，首先将具有更多样本的负类样本点划分为m个子集，以尽可能地保证每一个子集的样本点数目与正类样本的数目相当。接着，采用每一块负类样本子集与正类样本组合作为训练数据集进行训练支持向量机模型。最后将训练得到的各个模型应用到测试集上，得到各个相应的预测值（实值而非符号值），并对这些值取平均作为最终分类器预测得到的预测值，进而确定测试集的类别标号。具体流程图见图3-10。

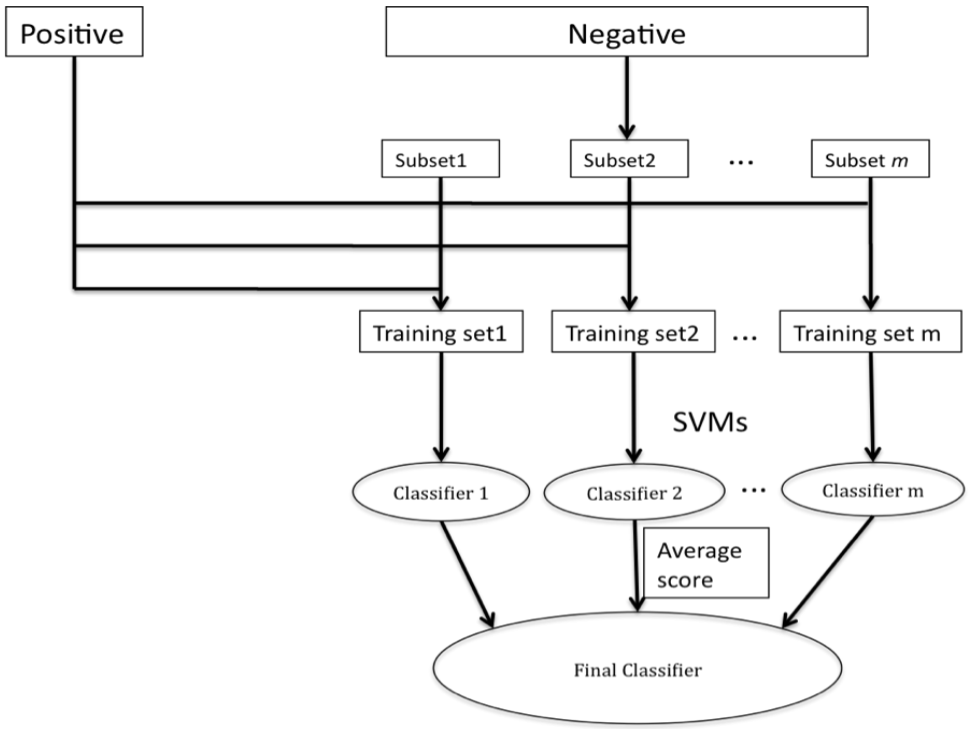


图3-10 不均衡数据集的处理策略

3.2.6 测试方法及评价指标

本节采用基于WW结构域的留一法（leave-one-domain-out）的测试方法进行测试分类器的性能，也即每次取出一个WW结构域（及与其相关的所有WW结构域-短肽对）作为测试集，利用其他剩余的样本作为训练集；如此反复，直至对所有的WW结构域全部经过测试。本节采用AUC作为两类分类器预测性能的评价指标。

3.3 WW结构域-短肽相互作用之两分类问题结果分析

3.3.1 不同特征编码方式的结果比较

以支持向量分类机为例，尝试不同的特征编码方式并比较它们之间的分类性能。用到的特征编码方式主要是二值正交编码、基于Profeat的特征编码、以及基于物理接触对的特征编码，包括

二值正交编码和成对氨基酸对之数值化编码。为了比较各种特征编码对应的性能，此处采用十折交叉验证进行测试。在对不同的特征编码方式进行选择最优参数之后，各自10次10折交叉验证的平均分类性能如表3-2所示。

表3-2 不同特征编码的支持向量分类机性能比较

Different encoding	AUC
Sparse20	0.948
Profeat+sparse20	0.88
Contact-pairs based sparse400	0.885
Contact-pairs based frequency	0.873

从上表3-2可以看出，虽然其他的编码方式也取得了相当的分类性能，但二值正交编码方式对应的分类器具有最佳的分类性能（AUC=0.948），所以本章后面的实验将采用二值正交编码方式作为最终的特征编码方式。

3.3.2 基于WW结构域的留一法

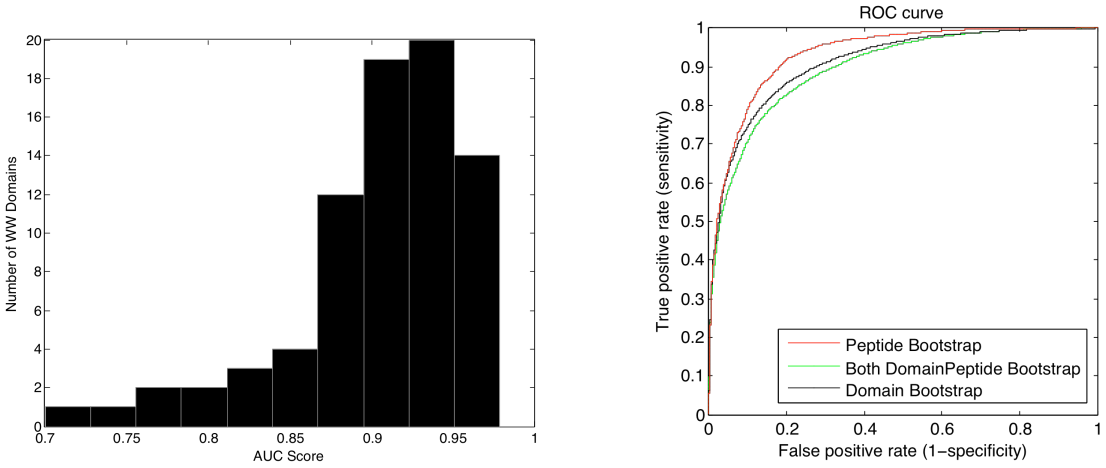
为了回答给定一个新的WW结构域，其结合（或者识别）短肽序列的特异性问题，我们采用基于WW结构域的留一法来模拟该问题。

首先，将特征编码后的WW结构域-短肽对作为支持向量分类机和支持向量回归机的特征输入，以正负类标号或者AU值作为两个模型的目标输入。

1. 支持向量分类机

通过基于WW结构域的留一法测试，基于二值正交编码和支持向量分类机，得到了针对每一个WW结构域的分类预测结果，并用AUC表示。对于78个WW结构域对应的AUC值，可以参见图3-11 (A)。

此外，为了更进一步准确估计支持向量分类机的平均分类性能及其相应方差，本节采用bootstrap测试法进行试验。主要从3个方面来进行评估，分别是对WW结构域进行bootstrap抽样测试、对短肽进行bootstrap抽样测试以及同时对WW结构域和短肽进行bootstrap抽样测试。这三种bootstrap的抽样测试分别模拟了“给定一个或者多个未知（未曾用以训练模型）的WW结构域，预测其与其他短肽是否相互作用的情形”、“给定一个或者多个未知的短肽序列，预测其与WW结构域是否相互作用的情形”以及“给定一个或者多个未知WW结构域，同时给定一个或者多个短肽序列，预测它们之间是否发生相互作用的情形”。我们重复了50次上述三种bootstrap测试，每次分别选取10%WW结构域的抽样测试、20%短肽序列的抽样测试、及其综合前面两者情形的抽样测试。具体分类器对应的平均预测性能如图3-11(B)。



（A）采用基于WW结构域的留一法测试得到的AUC值分布图 （B）采用3种不同策略的Bootstrap测试法得到的分类性能

图3-11 基于SVM的WW结构域-短肽相互作用对预测结果

综合上述结果，表明WW结构域-短肽之间的相互作用模式是可以直接从氨基酸的一级结构序列信息学到的。

2. 支持向量回归机

通过基于WW结构域的留一法测试，基于二值正交编码和支持向量回归机，得到了针对每一个WW结构域的分类预测结果，并用AUC表示。对于这78个WW结构域的AUC值分布图可参见图3-12。由图可见，利用SVR得到的模型，对WW结构域的分类性能也较佳。

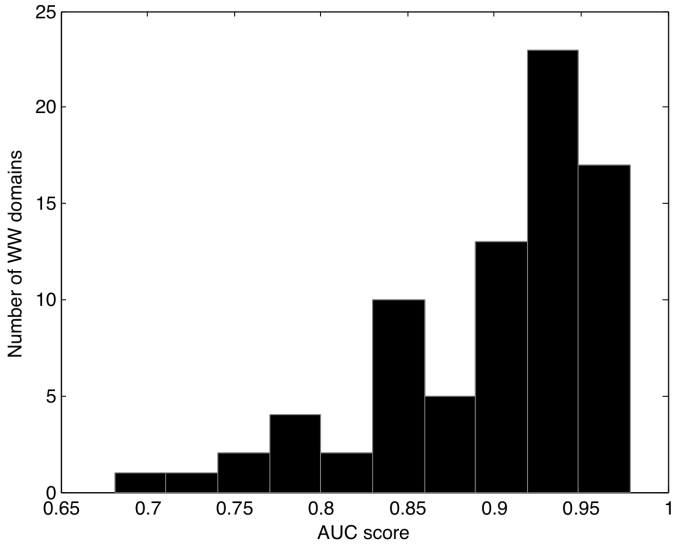


图3-12 基于SVR模型的WW结构域-短肽相互作用对预测结果

3. SVM 和SVR的结果比较

一个比较有趣而重要的问题是由实验测得的AU值对于构建预测WW结构域-短肽相互作用与否的模型是否有显著的效用。为此，比较了仅利用正负类信息的支持向量分类机模型和利用了AU值的支持向量回归机模型的分类性能，具体结果如图3-13。从该图中可以看出支持向量分类机的分类性能与支持向量回归机的分类性能基本相当，在单个WW结构域的分类性能上不分伯仲，均有优劣。这表明AU的值并没有给分类模型带来太多的额外信息。通过短肽序列对应的正负类信息结合支持向量分类机就可以较好地回答WW结构域-短肽相互作用与否的问题。所以本章将采用支持向量分类机作为最终的预测WW结构域-短肽相互作用与否的模型。基于此，我们建立了基于该模型的网络在线预测工具，以便对WW结构域感兴趣的生物学家进行预测。

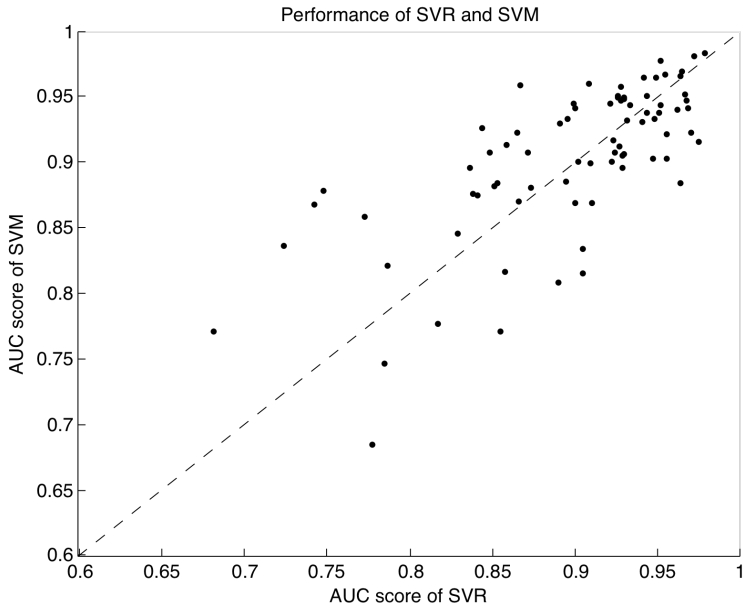


图3-13 SVM和SVR模型的结果比较

4. WW结构域-短肽相互作用预测网络工具介绍

本节对基于支持向量分类机的预测模型的网络工具进行简单的介绍。该预测工具使用方便，仅需要用户选定关注的WW结构域和输入所关注的短肽序列即可运行程序。其中网站提供了本章研究的78个WW结构域作为可选WW结构域，同时，用户既可以人为地输入一个或者多个12个长度的短肽序列（每一行对应一个12个长度的短肽序列），也可以通过文件的方式批量读入短肽序列（在文件中，也要求每一行对应一个12个长度的短肽序列）。此外，还可以实现对一个完整的蛋白质序列进行扫描（Scanning）预测是否含有相互作用短肽序列。具体的网络页面可参见图3-14，具体的网址为：http://www.baderlab.org/WWpredictor。

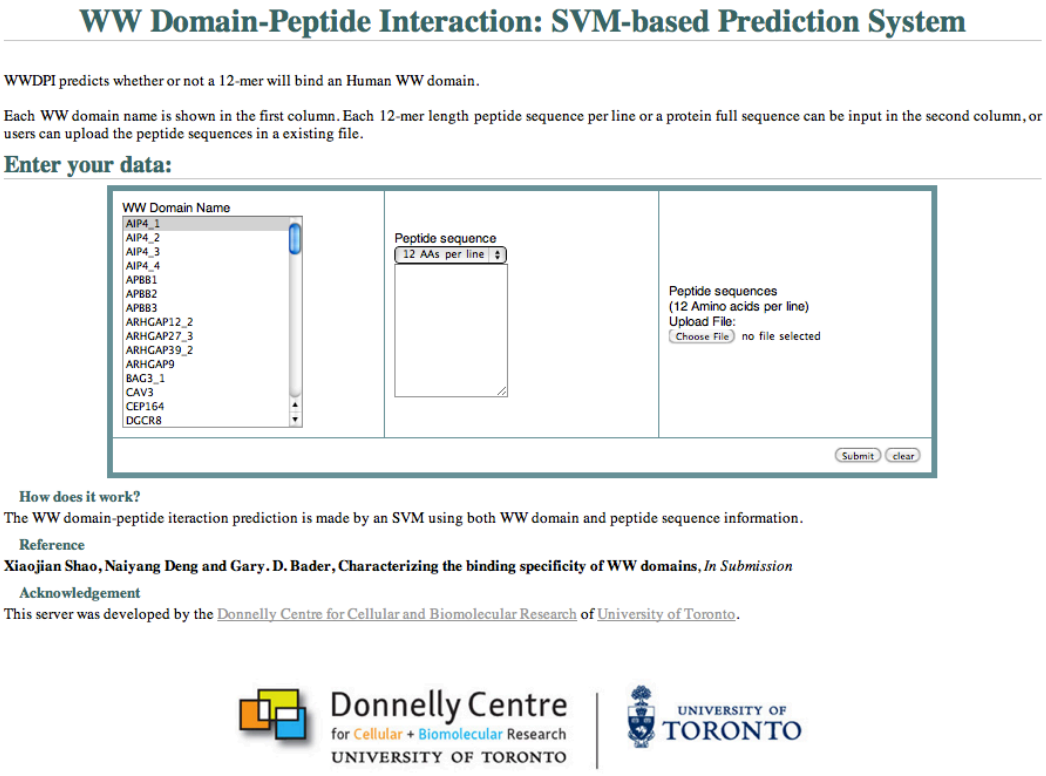


图3-14 WW结构域-短肽相互作用预测网络工具

对选定的WW结构域和输入的短肽序列预测完之后，该网络预测工具会返回一个预测结果页面，具体如图3-15所示。以“WWP1_1”为例，若输入短肽序列分别为“PGTAPPPYTVGP”和“FSNRPPGYPSQP”，则该工具分别预测为“正类”和“负类”，并给出了支持向量机模型的预测打分值。

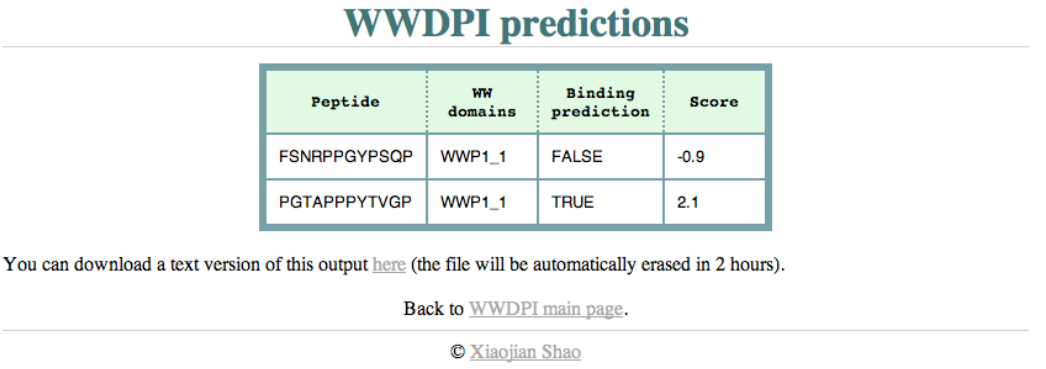


图3-15 WW结构域-短肽相互作用预测工具之预测结果图示

3.3.3 基于结构域家族的模型优于基于最近邻的模型

上面提到的支持向量分类机模型是基于整个WW结构域家族的模型，也即整合了所有除测试结构域之外的所有WW结构域及其相关的短肽相互作用对的信息来构建的模型。可以设想整合了其他结构域的信息有可能提升分类器的性能，同时也有可能会增加模型的噪音。为此，本节尝试比较了两种不同的构建分类模型策略。

基于最近邻分类模型：针对每一个测试WW结构域，在训练集中找到在序列上最近邻的WW结构域，然后基于该WW结构域-短肽相互作用对的数据学习出一个分类器，之后再利用该分类器去预测与测试WW结构域相关的所有相互作用对。

基于自身的分类模型：针对每一个测试WW结构域，不考虑其他WW结构域的任何信息，而仅仅考虑现有的关于该WW结构域的短肽相互作用对。利用常规的留一法进行交叉验证基于自身短肽相互作用对构建的分类模型的性能。

表3-3 预测WW结构域-短肽相互作用的不同策略的模型性能比较

方法	平均AUC
全局SVM	0.9
最近邻SVM	0.82
自身SVM	0.65

通过表3-3可以看出，基于“结构域家族”的分类器模型具有最佳的分类性能，这表明了该分类器能有效地整合各个不同的WW结构域及其相关的短肽相互作用对的信息。值得注意的一点是，事实上基于最近邻的分类模型也取得了较为不错的分类结果。这表明虽然基于结构域家族的分类模型取得了最佳的分类性能，但是大部分信息应该还是来自最近邻的结构域信息。另一方面，基于自身的分类模型并不能取得理想的分类性能，表明了仅从短肽序列的角度还不足以预测与给定WW结构域是否相互作用的问题。

3.3.4 不同的定义正负类的阈值

虽然本节一直利用0.5AU作为区别正类与负类（WW结构域与短肽是否相互作用）的阈值，但是到现在为止，我们并不确定选定其他不同的阈值（相当于从不同的角度定义WW结构域与短肽相互作用的强弱）是否会影响最终的分类器性能。为此，本节尝试了不同的阈值，分别选取了1AU、1.5AU、2.5AU和3.5AU作为定义正负类的阈值，并针对相应的数据集进行了分类器的建模。通过实验比较发现，支持向量分类机在各个阈值定义的分类问题上可以得到比较一致的，对应分类性能逐步提高的结果，相应的在78个WW结构域上的平均AUC分别为0.919、0.924、0.926和0.946。这表明如果需要识别（预测）更强（相比较0.5AU而言）的正类相互作用对，现有的特征编码和分类器方法能达到更佳的效果。

3.4 预测WW结构域-短肽相互作用强弱的多分类问题

前面提到，Hu等人得到的关于WW结构域-短肽相互作用的实验数据提供了具体的AU值，该值在一定程度上反映了WW结构域-短肽之间相互作用的强弱，并且根据AU值的分布可以定义出不同程度的正类相互作用对，例如较弱的相互作用对定义为亲和力介于0.5～1.5AU之间；弱相互作用对亲和力介于1.5～2.5AU之间；一般相互作用对亲和力介于2.5～3.5AU之间；强相互作用对定义为亲和力大于3.5AU。据此，得到了关于正类相互作用对的4大类别。本节希望利用机器学习方法，可以实现对任意一对正类WW结构域-短肽相互作用对的上述类别归属的预测。

由于本节关注的生物问题可以看成是普通的四分类问题，也即对于上面定义的{较弱、弱、一般、强}四类进行建模。另外一方面，注意到上述四类之间并不是简单的独立的类别关系，而是类与类之间其实存在着某种“序”上的大小或者相邻关系，也即各个类之间的样本对应的AU值的大小存在着由小往大递增的关系，例如，第2类与第1类和第3类相邻，但是第1类和第3类却不相邻。通常称此类问题为顺序回归问题。它是一个多分类问题，但是又比普通的多分类问题特殊，含有额外的类与类之间的序关系信息。此外，此处涉及的正类WW结构域-短肽相互作用对均对应有AU值，所以也可以考虑利用回归模型学出AU值之后再根据定义划分不同的类别归属。基于此，本节将分别采用支持向量多分类机模型、支持向量顺序回归机模型以及基于支持向量回归机的多类别模型来建模。本节首先介绍求解上述问题的若干机器学习算法，然后给出各个模型对应的分类性能比较。

3.4.1 本节用到的机器学习模型

1. 支持向量机多分类模型

求解多分类问题的支持向量机算法有很多，主要有基于多个两分类模型的“分而治之”策略和“整合模型”策略两大类。其中基于多个两分类模型的策略又主要分为：“1对1”和“1对多”两大类[44]。本节主要采用基于“1对1”策略的多分类支持向量机模型。利用LibSVM 软件实现，具体可以查阅Lin等人的工作[44]。

2. 支持向量顺序回归机模型

本节主要采用Chu等人发表的对基于shashua等人的支持向量顺序回归机的一个改进模型，如第一章所介绍的模型。在模型中用到的具体类别的划分如上节所讨论。

3. 支持向量回归机模型

本节采用标准支持向量回归机模型，具体在第一章已经有介绍。这里主要用到正类WW结构域-短肽相互作用对的AU值（即>0.5AU的数据）。并且利用LibSVM软件实现支持向量回归机模型。

3.4.2 预测WW结构域-短肽相互作用强弱的多分类模型

在前节根据相互作用对的AU值的具体分布，将正类WW结构域-短肽相互作用对分成了四类。具体的这四类中样本数的分布可以参见图3-16。

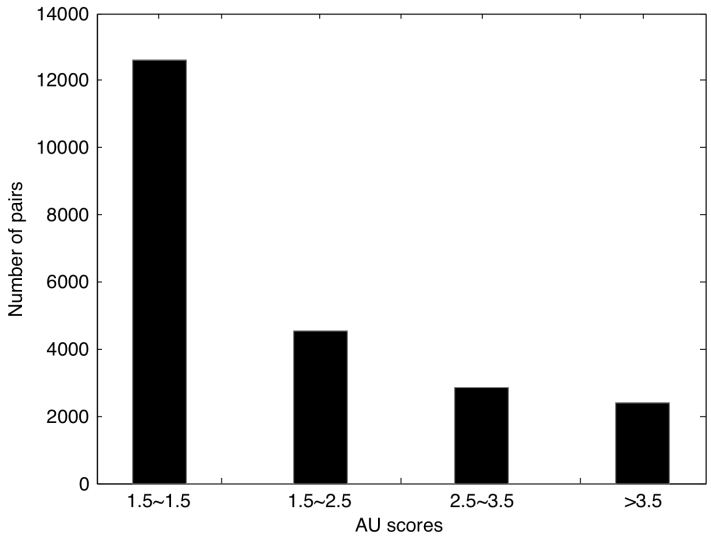


图3-16 WW结构域-短肽相互作用数据之正类集合中各个类别的样本数分布

基于上述样本类别的划分，利用公开软件Libsvm实现基于“1对1”策略的支持向量多分类机模型和支持向量回归机，采用文[48]中提到的SVORIM算法实现支持向量顺序回归机模型。在这三个模型中，均采用RBF核函数来实现。同时，基于十折交叉验证测试使用网格选参数的方法来选取模型中涉及到的参数的最优组合。基于最优参数，对上述三个模型进行了测试性能的评估。由于上面提及这四个类之间的某些类与类之间样本数相差大，在评估各个分类器性能的时候需要考虑这个因素。为此，除了采用通常的总分类精度这个指标之外，本节还细看了各个类的分类精度（灵敏度，sensitivity）。

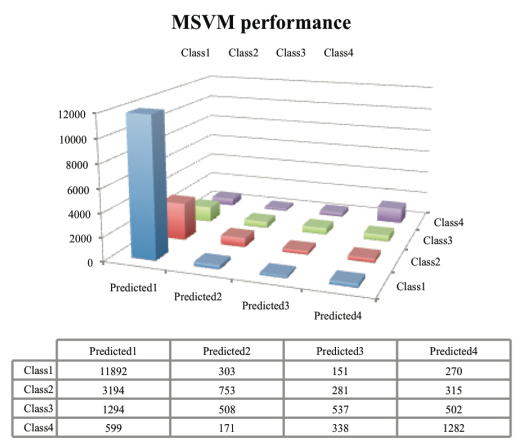
3.4.3 试验结果分析

本节主要展示上述模型在“正类”WW结构域-短肽相互作用对的类别归属预测性能。

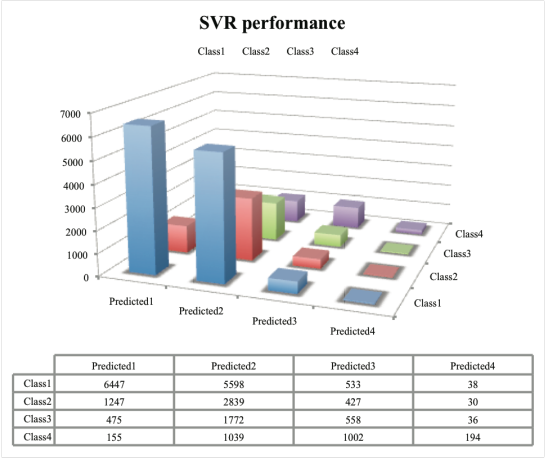
采用二值正交特征编码方式来表示WW结构域-短肽相互作用对，分别利用上述三种不同的模型来进行离散化描述相互作用强弱的多水平（也即多类别）的预测。我们也尝试了基于前述所用的其他特征编码方式，其预测性能并没有比二值正交特征编码方式好。所以此处只报道基于此特征编码的各模型之结果的比较。

基于十折交叉验证，具体的三个分类器的性能如图3-17所示，1对1多分类模型、顺序回归模型和支持向量回归机三个分类器的总分类精度分别为64.6%、65.4%和35.3%。具体到每一个类的分类正确率则如表3-4所示，三个分类器在4个相互作用水平上的平均分类精度分别为45.8%、56%和44.8%。

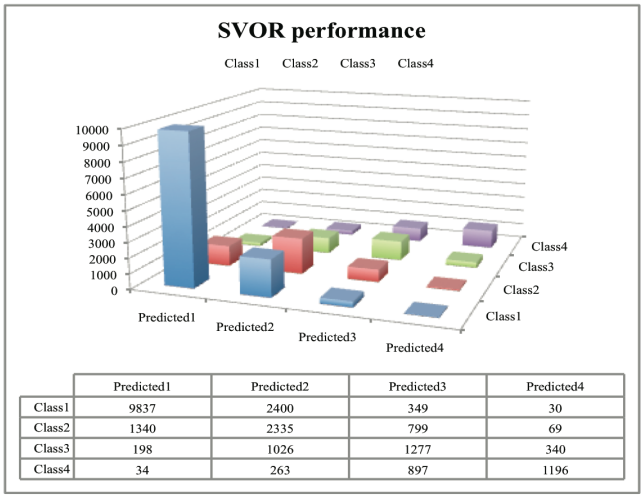
三个分类器中，顺序回归机分类性能最好，其次是1对1的多分类模型，最差的是支持向量回归机。这表明试验测得的AU值可能含有一定的噪音，尤其是在不同水平的临界值附近。顺序回归模型的性能比多分类模型要好，这也说明顺序回归模型有效的利用了不同类之间相互作用强弱（亲和力）水平的近邻信息，能较一般的1对1多分类模型更好地抓住类与类之间的顺序特性。



	Predicted1	Predicted2	Predicted3	Predicted4
Class1	11892	303	151	270
Class2	3194	753	281	315
Class3	1294	508	537	502
Class4	599	171	338	1282



	Predicted1	Predicted2	Predicted3	Predicted4
Class1	6447	5598	533	38
Class2	1247	2839	427	30
Class3	475	1772	558	36
Class4	155	1039	1002	194



	Predicted1	Predicted2	Predicted3	Predicted4
Class1	9837	2400	349	30
Class2	1340	2335	799	69
Class3	198	1026	1277	340
Class4	34	263	897	1196

图3-17 预测WW结构域-短肽相互作用强弱的多分类模型结果

MSVM代表的是采用1对1多分类支持向量机的模型结果；SVR代表采用支持向量回归机的模型结果；SVOR代表采用顺序回归支持向量分类机的模型结果。

表3-4 三种模型在各类上的分类灵敏度比较

Sensitivity	MSVM	SVR	SVOR
class1	94.26	51.10	77.97
class2	16.57	62.49	51.40
class3	18.90	19.64	44.95
class4	53.64	8.12	50.04

3.5 结论和工作展望

本章的工作实现了基于大规模实验数据训练得到的在整个人类蛋白质组层面上的针对WW结构域和其配体之间相互作用强弱（亲和力）预测的计算模型。本章的试验结果表明仅仅从WW结构域或者短肽配体的序列信息进行预测它们之间的相互作用是行之有效的，这给生物学家进行进一步的生物实验研究提供了非常重要和丰富的资源。

本章实验结果表明简单的基于氨基酸序列的二值正交编码方式可以用来表示任意一对WW结构域-短肽相互作用对，结合支持向量分类机等机器学习方法可以实现较精准地预测WW结构域-短肽之间的相互作用。实验结果也显示其他基于氨基酸物理、化学特性，基于蛋白结构模板的二值正交编码特征表示方式均未能比简单的基于序列的二值正交编码方式有更好的预测精度。这说明氨基酸序列信息本身富含了用于识别WW结构域-短肽相互作用的丰富信息。

通过仅利用正负类信息的支持向量分类机模型和利用了AU值信息的支持向量回归机模型之间的比较，发现利用了额外信息（AU值）的支持向量回归机模型并未能取得显著地好的预测性能，这也说明WW结构域和短肽序列信息本身富含了大量适合于预测是否相互作用的信息。也就是说，就回答WW结构域与其短肽是否相互作用的问题，相比于WW结构域和短肽的氨基酸序列信息，AU值似乎是冗余的。文章在定义WW结构域-短肽相互作用的正负类时，根据AU值的总体分布，采用了0.5AU作为阈值。鉴于不同生物学家对相互作用与否的认识不同，我们尝试了采用不同的阈值来定义正负类，并利用支持向量分类机进行分类建模，其结果表明如果定义具有更大的AU值的相互作用对为正类，支持向量机则能更容易地识别出来。这表明更强的相互作用对具有更显著的结合特性，更易于机器学习进行识别。

本章采用的机器学习策略是整合现有WW结构域家族的数据来构建分类器。这与传统的基于单个结构域建立分类器或者利用最近邻结构域信息构建分类器模型的策略不同。比较这三种不同的策略，发现基于单个结构域构建的分类器性能最差，这说明单单从短肽配体序列角度并不能有效地推广到其他未知的短肽。一种可能的原因是某些WW结构域可能会存在不同的结合模体，使得不同模体对应的短肽序列之间不能有效地相互传递信息（各个短肽之间氨基酸序列不太相似）。通过比较基于最近邻结构域的分类器和基于结构域家族的分类器，我们发现基于最近邻结构域的分类器性能虽然不及基于结构域家族的分类器，但是也能取得令人相对满意的分类精度。在生物学上，一般认为相似的结构域之间的结合短肽序列模体应该也是比较接近的，从这个角度来看，基于最近邻结构域的分类器能取得令人相对满意的结果也不足为奇。基于WW结构域家族的分类模型具有最佳的分类性能，这表明该模型能有效地整合不同WW结构域之间的短肽结合特性。换言之，该模型不仅能从最近邻的WW结构域数据集中学到知识，同时还能从其他非近邻的结构域中学到一定的知识。值得注意的是，整合较远的WW结构域信息有时可能会适得其反，可能会有增加噪音的风险。这一点在本章工作中并未得到证明，有待进一步的工作。未来可以尝试的策略是针对每个WW结构域，选取k个近邻的或者序列相似性高过一定程度的WW结构域的数据进行训练分类器模型，通过这种折中的策略，以期能充分利用WW结构域家族中其他结构域的信息，同时能尽量避免学习分类时受到噪音（例如，与测试WW结构域具有完全不同的结合短肽识别模体）的干扰。

本章用到的生物实验——蛋白质芯片技术得到的亲和力值（AU值）并不一定就是对应真正的生物体内发生的相互作用亲和力的大小，所以暂且不考虑利用其来预测WW结构域-短肽相互作用的亲和力。但是另一方面，该AU值与真正的亲和力是有正相关关系的（有可能是线性关系，也有可能是某种非线性保持单调的关系），在某种程度上体现了WW结构域-短肽相互作用的强弱（亲和力的水平）。基于此假定，我们利用新的机器学习方法——支持向量顺序回归机预测了WW结构域-短肽相互作用对的相互作用强弱的水平（亲和力水平）。试验结果表明本章采用的支持向

量顺序回归模型较好地模拟了不同亲和力水平之间的序关系的特性，与其他两个模型相比较，展示出了更高的预测分类性能。通过支持向量顺序回归机模型，可以进一步预测WW结构域-短肽相互作用对的相互作用强弱水平（亲和力水平），能给生物学家提供更细致的信息。

在蛋白质-蛋白质相互作用领域，很多生物实验得到的相互作用对之间的一致性很低。对于WW结构域-短肽相互作用问题，不同的实验手段也会得到非常不同的相互作用对。而在不同物种中进行的生物实验也往往会得到很多不同的相互作用对（哪怕是经过同源比对之后仍会有很多不一致的相互作用对）。如何有效地整合这些来自不同实验的数据集，甚至是来自不同物种的数据集，进而建立一个抗噪音的高性能分类器将是一个比较有趣的研究方向。

此外，如引言所述，有一部分WW结构域偏向于结合（或者识别）被磷酸化的（S/T）P短肽模体。例如文[87]中提到人类的PIN1中的WW结构域能与磷酸化后的短肽配体蛋白结合，但不能与未被磷酸化的同一短肽配体结合。然而，在同一文中他们也发现某些短肽序列只能在去磷酸化之后，才能与酵母ESS1蛋白中的WW结构域相结合，而一旦这些短肽序列被磷酸化，则相互作用将不再发生。类似的结果在文[137]中也有曾报道。他们发现配体保守短肽序列PPXY中的酪氨酸一旦被磷酸化，YAP的WW结构域将不能与之结合。上述例子说明这种WW结构域-短肽之间的相互作用是受磷酸化的负调控的。那么究竟在什么样的情况下，有些磷酸化是正调控，而有些磷酸化将是负调控的呢？在被磷酸化后负调控的情形下，是否因为有SH2结构域的协同或者竞争作用呢？这些问题的解答将有利于人们加深对WW结构域-磷酸化短肽相互作用的认识。而其中如何有效地识别哪些短肽配体是受磷酸化的负调控，对认识磷酸化影响调控WW结构域介导供体蛋白与短肽配体蛋白之间相互作用的机理起着至关重要的作用。

蛋白质-蛋白质相互作用是许多蛋白质行使功能的基础。越来越多的研究工作表明蛋白质-蛋白质之间的相互作用往往遵循一些特定的规律。主要是通过结构域与结构域之间的相互作用，或者特定结构域与短肽配体的相互作用，这两种特定的相互作用模式大概占据了90%左右的蛋白质-蛋白质相互作用[16]。相信随着蛋白质组学研究的不断完善和发展，生物技术的进一步提高，蛋白质-蛋白质之间相互作用的研究，尤其是结构域与短肽配体之间相互作用的研究也会进入更高的一个阶段，取得更大的成就。

第四章 PDZ结构域-短肽相互作用预测研究

第三章研究了WW结构域与其短肽之间的相互作用，本章将继续开展结构域与短肽之间的相互作用方面的研究，关注另一个非常重要的蛋白结构域——PDZ结构域。与第三章的关注点不同，此章关注定量化预测结构域与短肽之间的相互作用（也即回归问题），而第三章关注了WW结构域与短肽是否发生相互作用的分类问题以及它们之间相互作用的强弱的不同水平（基于顺序回归的多分类问题）。定量化测定结构域与短肽配体之间的相互作用大小（相互作用亲和力）可以更好地理解与该结构域发生相互作用的短肽蛋白之间的相互竞争关系，帮忙构建更具生物含义的蛋白相互作用网络，并且为药物靶点设计提供帮助。因此，定量化预测结构域与短肽配体之间的相互作用强弱具有非常重要的意义。本章以PDZ结构域为切入点，设计针对PDZ结构域及其短肽配体之间相互作用的定量化预测模型，可以实现从序列信息角度预测PDZ结构域与短肽之间的相互作用强弱。

4.1 引言

PDZ（PSD-95/Discs-large/ZO-1）结构域是另一种非常重要的信号传导的蛋白质结构域。PDZ结构域一般由80～100个氨基酸构成，主要含有2个α螺旋和6个β折叠。该结构域的名字源自最初发现含有该结构域（保守功能序列）的3个蛋白质：PSD-95（a 95 kDa protein involved in signaling at the post-synaptic density）、Discs-large（the Drosophila melanogaster Discs Large protein）、和ZO-1（the zonula occludens 1 protein involved in maintenance of epothelial polarity）的首字母的缩写。PDZ结构域又称GLGF（因为最初发现的PDZ结构域含有以GLGF开头的保守序列）或者DHR结构域（因为含有discs larget同源区域）。PDZ结构域的图示见图4-1。该结构域是近年来最为常见的蛋白质相互作用结构域，广泛存在于各种生物体内。根据Pfam数据库[6]（版本号24.0，PDZ标号PF00595）最新的统计，后生生物中线虫含有184种，果蝇有246种，人类则超过500多种；但是非常奇怪的是很少存在于酵母中（只有2-3种）。这分别对应了92个线虫蛋白质、170个果蝇蛋白质、超过400多个人类蛋白质；而只有3个*S.pombe*蛋白质和2个酿酒酵母蛋白质[138]。由此可见，一个含有PDZ的蛋白质可以含有一个或者多个PDZ结构域，其中部分含有PDZ结构域的蛋白质结构域构架图可参见图4-2。事实上，含有PDZ结构域的蛋白质也还可以含有其他结构域如SH3、LIM和WW结构域等串联，一块完成信号传导、蛋白质复合体装配等特定生物功能。

PDZ结构域主要起到调控信号传导、铁离子通路、细胞极性化、神经元发育、装配分子信号传导蛋白质复合物等作用，与很多人类疾病相关，例如，神经系统发育、精神分裂症等[139]。PDZ结构域主要是通过与其他短肽配体相互作用来行使功能，主要识别配体蛋白质的C端序列。早期的研究表明，根据C端短肽序列的两个氨基酸位置的氨基酸类别不同，PDZ结构域的短肽结合模式主要有3大类：X[T/S]XΦCOOH、XΦXΦCOOH和X[ED]XΦCOOH，其中“X”为任意一个氨基酸，Φ为亲水氨基酸（hydrophobe），[S/T]表示该位置要么为S，要么是T。其中前两类是最为常见的。

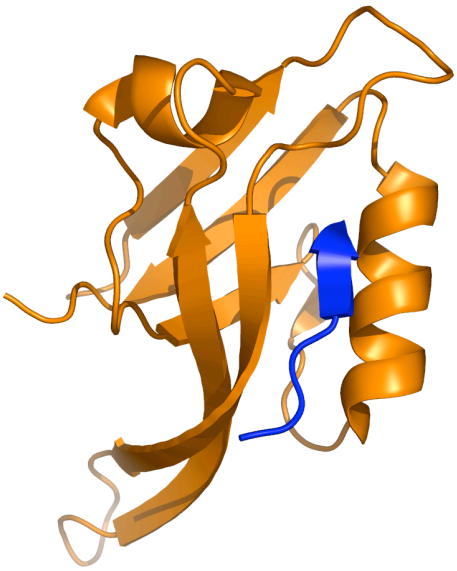


图4-1 PDZ结构域与短肽C-末端结合三维结构示意图

图中为Par-6 PDZ结构域和VKESLV-COOH短肽复合物，PDB数据库中ID号为1RZX。

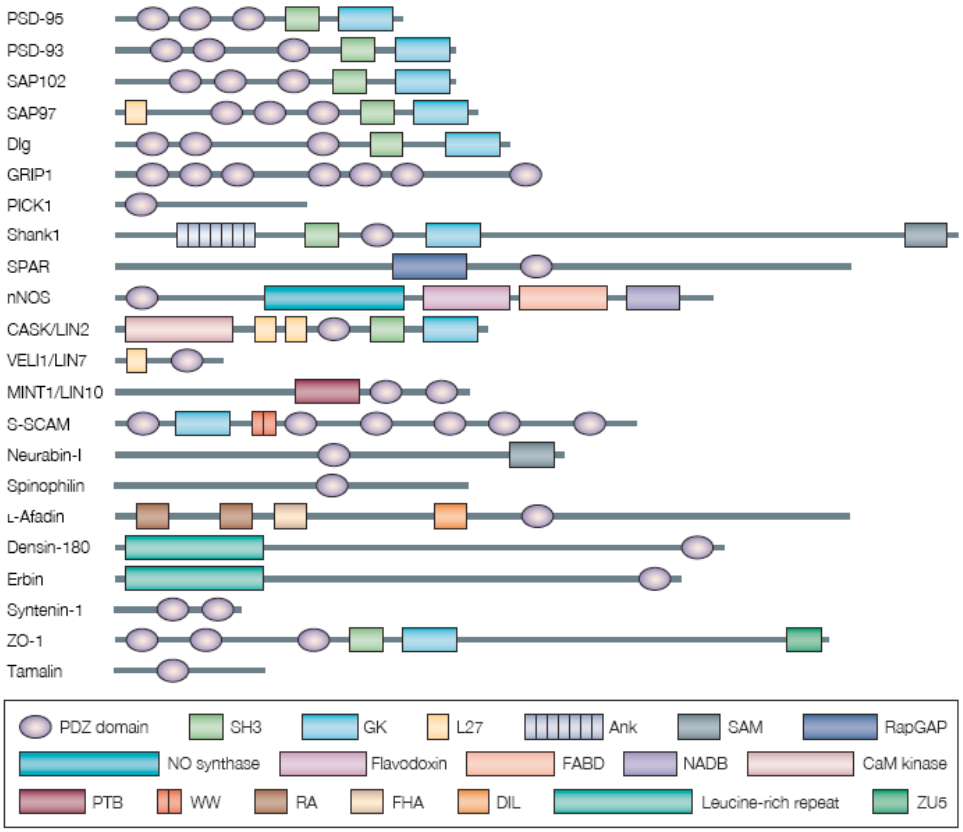


图4-2 含有PDZ结构域蛋白质的结构域架构图

某些蛋白质只含有一个PDZ结构域，而有些蛋白则可以含有多个PDZ结构域。该图摘自文献[140]。

在前人的研究工作中，绝大部分集中在研究PDZ结构域的结合特异性问题上。研究蛋白质结构域的结合特异性可以帮助理解蛋白质结构域如何实现与特定短肽之间的相互作用，找到特定的结合规律，并根据此规律发现新的蛋白质-蛋白质相互作用对，同时还可以帮助指导短肽的设计和PDZ结构域的合成。正是由于该生物问题的重要性，许多学者分别用实验和计算的方法对PDZ结构域的结合特异性进行了研究。目前高通量实现PDZ结构域-短肽相互作用对的实验手段主要有蛋白质芯片技术[94, 95, 141]和噬菌体展示技术[92]等。基于上述高通量得到的数据集，这些小组开发了不同的计算模型。Tonikian等人基于噬菌体展示技术得到的数据设计了位置特异性打分矩阵（PWM）模型并预测了人类PDZ结构域的结合特异性[92]。Stiffler等人提出了一种基于多结构域选择性计算模型（multidomain selectivity model, MDSM），考虑了不同的PDZ结构域在短肽序列不同位置上对各个氨基酸的偏好程度，是对传统位置特异性矩阵的一种变形[94]。而Chen等人则结合了多个PDZ结构域的数据，提出了一种新的基于贝叶斯思想的Additive模型，利用该模型实现了对整个PDZ结构域家族的结合特异性的预测[95]。此外，Eo等人提出了基于支持向量机的预测模型，实现了对G耦合（G coupled）蛋白质中的PDZ结构域的结合特异性的测定[142]。最新的研究成果是Hui等人有效地整合了噬菌体展示数据和蛋白质芯片数据，结合支持向量机模型，提出了一种新的预测模型，并预测出了部分不同的PDZ结构域在人类蛋白质组、线虫和果蝇蛋白质组中的潜在结合配体，为未来进一步的实验提供了帮助[93]。

研究蛋白质-蛋白质之间的相互作用之结构域-短肽相互作用时，除了关心结构域的结合特异性问题之外，另一个非常重要的问题是回答结构域与短肽配体之间的相互作用强弱。通常短肽识别模块（结构域）与其短肽配体发生相互作用时的强弱范围大致是在电离系数为10*u*M的弱相互作用范围内。为了更深入地理解生物体内蛋白质-蛋白质相互作用的网络，不仅需要知道结构域会与哪些短肽配体发生相互作用，更需要预测出具体的结构域与其短肽之间的相互作用强弱（相互作用亲和力）。因为只有知道了结构域-短肽之间的相互作用的强弱，才能更好地理解与该结构域发生相互作用的短肽蛋白质之间的相互竞争关系，同时也能为人们设计药物靶点提供必要的帮助。事实上，知道了相互作用的强弱也能帮助人们更深入地理解生物体内许多调控网络，例如由转录因子调控的各个靶点基因之间被激活是有次序的，而要知道这种先后顺序就需要获知转录因子与各个靶点基因之间的相互作用亲和力的大小[143]；在蛋白质后翻译修饰的过程中磷酸化起了关键重要，Lew等人指出了FGFR1的磷酸化位点被磷酸化的过程也是有序的，该有序的过程受到了相互作用亲和力大小的影响[144]；在酿酒酵母中，High Osmolarity Glycerol（HOG）反应通路能对外界的高渗透压（osmolarity）的压力变化产生快速的反应，而这种反应与蛋白质之间的相互作用强弱有非常大的关系，也就是说，知道了Sho1p蛋白质中的SH3结构域与Pbs2p短肽之间相互作用亲和力的变化就能更好地理解HOG反应通路是如何对外界环境的变化做出相应反应的[145, 146]。

目前，在整个基因组层面上知道亲和力大小的由短肽识别模块（结构域）介导的结构域-短肽相互作用数据还不多。其中，主要组织相容性复合体（Major histocompatibility complex，MHC）作为一种免疫系统的结构域，在其与短肽配体之间相互作用定量化（亲和力）方面已经得到广泛的研究，并已产生大量的定量化数据[147-150]。从计算模型的角度来预测MHC以及相应的短肽配体之间的相互作用强弱的工作也已有诸多报道。其中，有针对单个MHC分析结构域构建的模型。

Liu等人针对小鼠中3个第一类MHC分子结构域，设计了11种生物化学特性进行表征各个氨基酸，并通过构建支持向量回归机模型（SVRMHC），分别实现了这3个MHC I分子结构域与相应短肽配体之间的相互作用亲和力的预测[85]。同时，也有整合整个MHC结构域家族数据的预测模型。Nielsen等人针对人类第一类MHC分子结构域，通过整合全MHC家族的数据，构建神经网络模型（NetMHCpan），实现了其与相应短肽配体之间的相互作用亲和力的预测[86]。之后，Nielsen等人又针对人类第二类MHC分子结构域，设计了基于整个MHC II结构域家族的神经网络模型（NetMHCIIpan）[151]。上述两个模型都整合了结构信息，通过考虑与短肽配体发生物理接触的WW结构域结合位点和短肽配体序列信息进行预测。在前人的工作中，针对MHC I类的预测模型已经比较完善并取得了较佳的预测性能，而对于MHC II类的预测模型，则由于受限于目前有限的MHC II类分子数据进行训练，使得其模型预测适用性大大减弱。虽然MHC是基于免疫系统的结构域，还不是传统意义上的短肽识别模块，但是基于MHC I或者MHC II类分子结构域与短肽相互作用的预测工作可以为研究信号传导的结构域的相互作用提供启示性作用，值得学习借鉴。

PDZ结构域与短肽相互作用的亲和力数据是目前第一个在整个蛋白质组层面上得到的在信号传导过程中涉及的短肽识别模块的数据[94]。该数据集是对小鼠蛋白质组中85个PDZ结构域和217个短肽蛋白质进行实验之后得到的大规模数据集，该数据集不仅测得了具有相互作用的PDZ结构域-短肽对之间的亲和力（电离常数K_d值）（正类样本集，定量化），而且还得到了部分确认未发生相互作用的PDZ结构域-短肽对构成的数据集（负类样本集，定性化；对于该数据集该实验未能对结构域-短肽对测得具体的K_d值，而是仅仅知道K_d值超过100uM），本章称这种类型的数据为“半量化”数据集。本章的工作基于这组大规模的数据集进行建模，从计算的角度利用蛋白质序列的信息实现对未知PDZ结构域和短肽之间的相互作用强弱的预测。根据上述数据的半量化特点，设计了一种新的支持向量回归机模型——半量化支持向量回归机，该模型能同时利用定量化的正类样本集和定性化的负类样本集信息。首先，在小鼠蛋白质组数据集上进行新算法的交叉验证。其次，利用新模型对发生单点变异的短肽序列进行预测来验证算法的有效性。此外，尝试利用新模型预测人类蛋白质组中的PDZ结构域-短肽相互作用强弱，并进行了实验验证。上述试验的结果均表明本章提出的新模型能有效地预测PDZ结构域与蛋白质短肽之间的相互作用强弱。

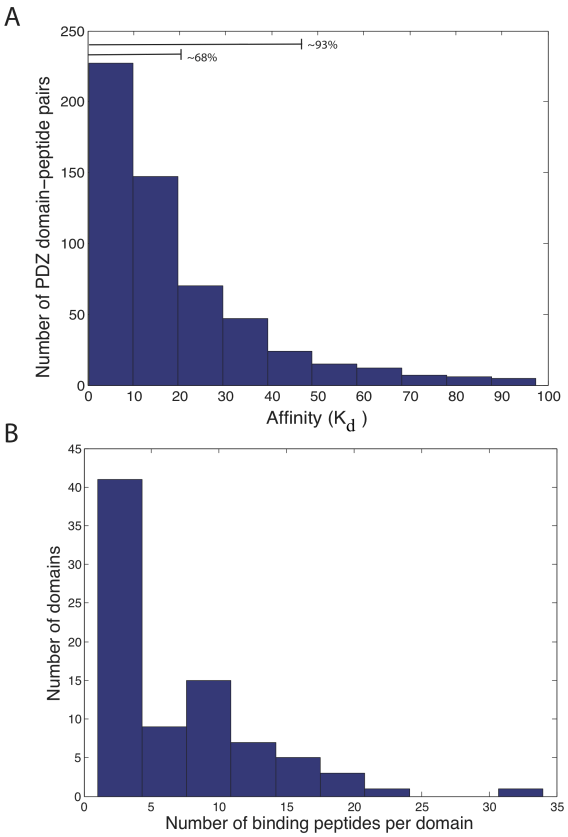
4.2 数据集和方法

4.2.1 数据集

本章用来建模分析的数据集来自文[95]，该数据集最初来自文[94]，前者数据集是对后者数据集的一个更新。该数据集涉及了小鼠蛋白质组中的85个PDZ结构域和217个蛋白质的C端短肽序列。该实验首先利用蛋白质芯片技术（peptide microarray）初步测得PDZ结构域-短肽相互作用对；然后利用一种高质量的测相互作用亲和力的技术——荧光偏振技术（fluorescence polarization，FP）测定相互作用对的亲和力的大小（电离常数值），具体流程参见文献[94]。在文[95]中最后实验测得了560对正类PDZ结构域-短肽相互作用对，涵盖了82个PDZ结构域和93个短肽序列；也测得了1167对负类PDZ结构域-短肽相互作用对，涵盖了82个PDZ结构域和138个短肽序列。该实验还得到了蛋白质芯片技术测得的负类相互作用对，只是这些负类相互作用对并未得到荧光偏振技术的

测定。

在该数据集中，560对正类PDZ结构域-短肽相互作用对对应的亲和力大小（电离常数，K_d值）均小于100*u*M（其中K_d值越小，对应相互作用亲和力越强；K_d值越大，亲和力越小）。并且事实上其中大部分的正类相互作用对的K_d值小于50*u*M（约占93%），约占68%的相互作用对的K_d值甚至小于20*u*M，具体亲和力K_d值的分布见图4-3(A)。如前所述，该实验得到的1167对PDZ结构域-短肽负类相互作用对，并未测得具体的K_d值，而仅仅知道它们的K_d值是大于100*u*M的。通过统计发现，在这82个PDZ结构域中，含有10个以上的结合短肽的PDZ结构域仅有23个。该数据的相关信息可以参见图4-3(B)，更为具体的数据信息可从http://Baderlab.org/PDZAffinity下载。



(A) PDZ结构域-短肽相互作用对的亲和力分布 （B）单个PDZ结构域的结合短肽数分布

图4-3 PDZ结构域-短肽相互作用数据统计

4.2.2 方法——预测模型

本章预测PDZ结构域-短肽相互作用强弱的问题属于回归问题的范畴。而一般的求解回归问题的方法，如线性回归模型、支持向量回归机等只能处理因变量值为连续取值的数据。但是由于本章涉及的数据的形式比较特殊，同时含有带有亲和力为连续取值的正类样本和仅知亲和力大于100*u*M的定性取值的负类样本。我们考虑设计新的适合这类数据形式的回归模型。本章设计了基

于支持向量回归机的新回归模型——半量化支持向量回归机，这里加以详细介绍。

给定数据集 $S=\{(x_i,y_i): x_i\in R^n, y_i\in R\}_{i=1}^m \cup \{z_j : z_j\in R^n\}_{j=1}^k$，其中$x_i$, z_j分别对应为正类PDZ结构域-短肽相互作用对和负类相互作用对，且均为n维的向量；y_i对应为正类PDZ结构域-短肽对x_i的相互作用强弱的亲和力（电离系数K_d值）， 同时已知负类PDZ结构域-短肽相互作用对z_j的亲和力（K_d值）大于给定的阈值（例如，$\hat{y}$ = 100uM）（可以定义其为先验知识）。目标是通过这种类型的数据求解一个回归函数：

$$f(x)=\sum_{i=1}^{m}\alpha_i K(x,x_i)+b, \tag{4-1}$$

以便可以利用$y=f(x)$来推断PDZ结构域与短肽相互作用的亲和力。并希望$f(x)$满足如下限制：

针对正类PDZ结构域-短肽相互作用对x_i而言，希望最小化基于“ε-insensitive”损失函数的预测模型对应的回归误差值$|f(x_i)-y_i|\le\varepsilon$，也就是说在训练集$x_i$上对应的回归函数值 $f(x_i)$的误差应该小于等于ε, i = 1, 2, …, m。而对于负类PDZ结构域-短肽相互作用对z_j而言，希望其对应的回归函数值能满足先前给定的知识信息（即回归函数值大于给定的阈值 $\hat{y}$ = 100uM）：$|f(z_j)-\bar{y}_j|\le\varepsilon$ 和 $\bar{y}_j\ge\hat{y}, j=1, 2, \ldots, k$。其中回归函数$f(x)$中的$x_i$为正类PDZ结构域-短肽相互作用对，$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_m)^T$ （T为转置符号，m为训练集样本规模）是类似于标准支持向量回归机中的Lagrange乘子，b是回归函数的偏置阈值；$K(x, y)$为支持向量回归机模型中用到的核函数。

上述提到的约束要求最终得到的回归函数$f(x)$在训练集S上的预测误差必须是在ε范围之内。但是这种约束显然太过于严格，在现实问题中，这些约束往往不能得到完全的满足。所以如标准支持向量回归机一样，希望引入一些松弛变量ξ_i和ξ_j，使得上述约束能尽量地得到满足。综合以上各种约束，类似Mangasarian等人提出的基于知识的回归模型[152, 153]，可以给出如下具有线性规划形式的半量化支持向量回归机模型：

$$\min_{\alpha,b,\bar{Y},\xi,\bar{\xi}} \ \|\alpha\|_1+C_1\sum_{i=1}^{m}\xi_i+C_2\sum_{j=1}^{k}\bar{\xi}_j \tag{4-2}$$

$$s.t. \quad \sum_{i=1}^{m}\alpha_i K(x_i,x_i)+b-y_i-\xi_i\le\varepsilon, \tag{4-3}$$

$$y_i-\sum_{i=1}^{m}\alpha_i K(x_i,x_i)-b-\xi_i\le\varepsilon, \tag{4-4}$$

$$\sum_{i=1}^{m}\alpha_i K(x_i,z_j)+b-\bar{y}_j-\bar{\xi}_j\le\varepsilon, \tag{4-5}$$

$$\bar{y}_j-\sum_{i=1}^{m}\alpha_i K(x_i,z_j)-b-\bar{\xi}_j\le\varepsilon, \tag{4-6}$$

$$\bar{y}_j\ge\hat{y}, \tag{4-7}$$

$$\xi_i\ge 0, \bar{\xi}_j\ge 0, \tag{4-8}$$

$$i = 1,2,\cdots,m;\ j = 1,2,\cdots,k, \tag{4-9}$$

其中ε是事先取定的定义损失函数的参数，C_1，C_2是惩罚参数，起着权衡正则项（避免过拟合）和经验风险误差（根据ε-损失函数定义，如下文）的作用。约束(4-3) - (4-4)保证在允许误差范围内以ε-精度使正类训练样本点预测正确；约束(4-5) - (4-7)则确保负类训练样本点在允许误差范围内以ε-精度满足先验知识。在该模型中，采用了1-范数度量误差项，也采用1-范数来度量模型的复杂度或者稳定性。这么做的其中一个原因是为了保持模型具有线性规划形式，比采用2-范数对应的二次规划模型要简单很多。此外，先前的很多工作均表明采用1-范数度量正则项能达到与2-范数相当的效果[154]。上述模型中用到了标准支持向量回归机中的ε-不敏感损失函数，该函数定义参见第一章内容。

半量化支持向量回归机的输入向量为经过特征编码之后的PDZ结构域-短肽相互作用对（见下文特征编码部分）以及对应的相互作用亲和力的K_d值；输出值则是相互作用的亲和力的预测值（越高的预测值代表了越弱的相互作用，越低的预测值则代表越强相互作用）。此外，本章涉及到的模型含有对回归性能有显著影响的诸多参数，如惩罚参数C_1，C_2和高斯核参数 σ 等。这里采用网格搜索方式进行对最优参数的搜寻。为了计算模型方便，本章对回归模型中用到的相互作用亲和力均进行数据归一化，也即将K_d值取log10之后再归一化到[-1,1]。半量化支持向量回归机模型是通过Matlab2008编程实现的，具体源程序可从http://baderlab.org/PDZAffinity下载。

为了与本章最终采用的模型相比较，本章还考虑了基于最近邻的半量化支持向量回归机模型。对于每一个待测试的PDZ结构域及其短肽对，选择离它最近的PDZ结构域及其短肽序列作为训练点训练半量化支持向量回归机模型；并且要求选择的最近邻PDZ结构域有超过10个结合短肽序列，以保证有充分多的训练点（也即若序列最近邻PDZ结构域对应的结合短肽个数不超过10个，就找次近邻的，依次类推，直至找到含有10个或以上结合短肽的“最近邻”的PDZ结构域）。

通常而言，针对短肽识别模块介导的相互作用预测模型主要有两大类：一类是基于单个结构域以及对应的短肽数据构建预测模型（记为“单结构域”模型），如针对单个结构域的基于短肽的PSSM或者PWM预测模型；另一类是基于多个结构域的模型（记为“多结构域”模型），可以有效地整合不同PDZ结构域以及相应的短肽相互作用数据，如本章提出的半量化支持向量回归机模型。

4.2.3 特征编码

通常在机器学习之前，需要对蛋白质序列进行特征编码。与WW结构域-短肽相互作用对的编码方式类似，可定义如下两大类编码方式：可以对PDZ结构域和短肽两个部分各自分别进行编码，然后组合起来；也可以直接从PDZ结构域-短肽对入手从整体上设计编码方式。

针对第一类编码方式，在前章中用来描述WW结构域-短肽相互作用对中WW结构域序列的均可以用来表示PDZ结构域-短肽相互作用对中的PDZ结构域序列。本章采用了二值正交编码方式和基于物理化学特性的Profeat编码方式[112]。此外还采用了对20个氨基酸根据不同物理、生化特性进行数值化的Zscale[155]、5factor[156]、11factor[85]等编码方式。除了Profeat之外（因为短肽序列的长度太短），其他方法也均可以用来表示短肽序列。

针对第二类编码方式，本章尝试了由PDZ结构域序列和短肽序列构成的各种不同的组合对，并采用二值正交编码方式来表示PDZ结构域-短肽相互作用对。这其中包括由PDZ结构域全长序列（这里采用根据Pfam数据库定义的多序列比对之后的118个氨基酸长度序列为PDZ结构域的全长序列）与短肽全长序列之间两两位置的结合对，包括PDZ结构域中的结合位点（根据a1-synPDZ结构域的三级结构信息得到的PDZ结构域上的16个结合位点或者经由9个不同PDZ结构域中均出现的10个核心结合位点）与短肽全长序列之间两两位置构成的结合对。以上提到的结合位点是根据通过X-结晶技术得到的蛋白质三维结构信息确定的，其中16个结合位点是由Chen等人在文[95]中提出的，而10个核心结合位点则是由Tonikin等人在文[92]中通过比较9个含有PDZ结构域和短肽的蛋白质复合体三维结构确定的。值得注意的是，这种两两之间的氨基酸对构成的特征编码向量是非常高维的。例如，由PDZ的16个结合位点和10个长度的短肽序列之间的两两之间的氨基酸对共有16×10 = 160对，如果每个对均采用二值正交编码（20×20，涵盖所有氨基酸对的组合可能）表示，则最终的特征向量为160×400 = 64,000维稀疏向量。为了降低编码维数（或者说简化编码计算），定义了一种结合“多项式核函数”特性的新的编码方式来达到类似的效果。设PDZ结构域和短肽序列可以表示为$PDZ = (P_1, P_2, ..., P_n)$，$Peptide = (pep_1, pep_2, ..., pep_k)$，其中$P_i$和$pep_j$分别代表PDZ结构域序列中第$i$个位置的氨基酸和短肽序列中第$j$个位置的氨基酸，并且此处对应$n = 118$，$k = 10$。通常PDZ结构域-短肽相互作用对就可以利用它们之间的“外积”进行表示：$PDZ*peptide = (P_1pep_1, P_1pep_2, ..., P_1pep_k, P_2pep_1, ..., P_npep_k)$（此处对应了$n \times k = 1180$个氨基酸对，如果利用前述的二值正交编码的话，就会得到1180×400=472,000维稀疏向量，这在实际应用中是相当巨大的维数，会大大影响之后的机器学习训练过程）。原则上，在采用支持向量回归机或者其他基于核方法的机器学习模型时，任意常用的核函数均可用来度量任意两个上述定义稀疏向量之间的相似性。幸运的是，任意两个由外积得到的向量之间的内积（即一个外积向量对应一个结构域-短肽相互作用对）可以表示为两个内积之间的乘积。例如：任意两个PDZ结构域-短肽相互作用对，设$(PDZ_1, peptide_1)$ 和 $(PDZ_2, peptide_2)$，它们各自之间的外积向量（即$PDZ_1*peptide_1$和$PDZ_2*peptide_2$）之间的内积可以表示为两个内积之间的乘积：$(PDZ_1^TPDZ_2) \times (peptide_1^Tpeptide_2)$。这样就不需要计算用正交二值编码表示的PDZ结构域-短肽相互作用对对应的472,200维稀疏向量之间的内积，而仅仅需要计算维数相对较低的两个PDZ结构域之间的内积或者相应的两个短肽序列之间的内积，从而大大降低了计算复杂度。

至此，将度量任意两个PDZ结构域-短肽相互作用对之间的相似性的过程转换为求任意两个PDZ结构域之间和任意两个短肽序列之间的相似性。利用核技巧，可以将两个向量之间的内积拓展到两个向量之间的核函数。进而，度量两个外积向量之间的内积（核函数）最后可以转换为两个核函数之间的乘积（一个核函数对应度量PDZ结构域之间的相似性，一个核函数对应为度量短肽序列之间的相似性）。本章采用“多项式核函数”来度量任意两个PDZ结构域或者短肽序列之间的相似性。利用该核函数，蛋白质序列可以直接用来作为核函数的输入向量，而不需要有特征编码的过程，所以既方便又不会产生巨高维的向量。此处定义“多项式核函数”如下：$K_{poly}(x, y) = (K_{baseline}(x, y)+1)^p$，其中$x,y$可以同时为PDZ结构域序列，或者同时为短肽序列；$p$为多项式对应的阶数。$K_{baseline}(x, y)$度量输入序列$x, y$在相同位置上具有相同氨基酸的位置总数。由于PDZ结构域序列与短肽序列具有不同长度的氨基酸序列，对应的$K_{poly}(x, y)$会因序列长度不同而又很大的区别，

所以需要对得到的$K_{poly}(x, y)$进行归一化，具体方式如下：$K_{poly\_normalize}(x,y) = K_{poly}(x, y)/\sqrt{K_{poly}(x,x) \times K_{poly}(y,y)}$。

为了便于与前人工作的比较，本章也尝试了Chen等人提出的关于PDZ结构域-短肽相互作用对的38个氨基酸物理接触对（图4-4(A)）编码。该信息是通过a1-synPDZ结构域与短肽序列GVKESLV复合而成的三级结构信息得到的，涉及到PDZ结构域的16个结合位点和短肽序列的最后C端的5个结合位点（参见文[95]）。对这38个氨基酸物理接触对采用正交二值编码方式进行特征编码，最终得到了38×400 = 15200维的高维向量。本章对不同核函数进行比较之后发现其他核函数并未得到比线性核函数好的预测性能，所以对于半量化支持向量回归机而言，针对基于38个氨基酸物理接触对的正交二值特征编码向量，采用线性核函数进行建模。

以上特征编码方式中，除了基于Profeat的编码方式外，其他的特征编码方式均需要对PDZ结构域序列进行多序列比对，以便确保原本序列长度不同的PDZ结构域具有相同维数的特征。本章采用文[95]中的PDZ结构域多序列比对的结果，并且根据Pfam数据库定义多序列比对之后其中保守的序列（经多序列比对之后的第22个位置到第139个位置共118个氨基酸长度的序列）为PDZ结构域序列，具体结果可以参见图4-4(B)。对于短肽序列，本章则采用10个氨基酸长度序列。

本章经过对不同特征编码以及不同核函数对应的半量化支持向量回归机的预测性能进行比较之后，发现“多项式核函数”对应的结果最好，其中针对PDZ结构域和短肽序列的相应的多项式阶数均为2阶。

4.2.4 预测性能度量指标

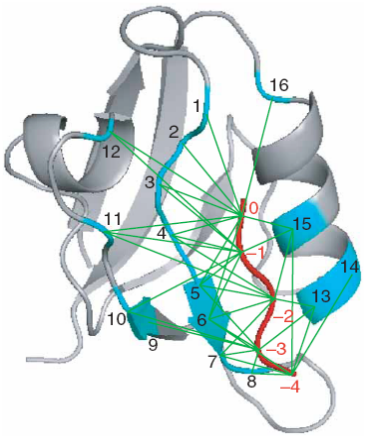
通常计算模型的预测性能最终需要通过实验方法来进行验证。但是众所周知，实验的手段是既费时又费钱的。为此，采用基于PDZ结构域的留一法来模拟并评估模型的预测能力。基于PDZ结构域的留一法即指每次测试时，取出与测试PDZ结构域相关的所有PDZ结构域-短肽相互作用对，训练点中不含有该测试PDZ结构域的任何信息（这个与实际情况更为相符）。也就是说每次选取一个PDZ结构域及其相应的短肽相互作用对（包括正类相互作用对和负类相互作用对）作为测试集，将余下的所有的PDZ结构域-短肽相互作用对作为训练集。对所有的PDZ结构域重复上述过程。

为了评估模型的预测能力，采用Pearson相关系数去评估模型预测亲和力大小（线性关系）的预测能力；采用Spearman相关系数去评估模型预测亲和力相对大小（有可能是非线性关系）的预测能力。本章采用Matlab中的“corr”函数来评估模型的模型预测能力（即计算出各种相关系数）。

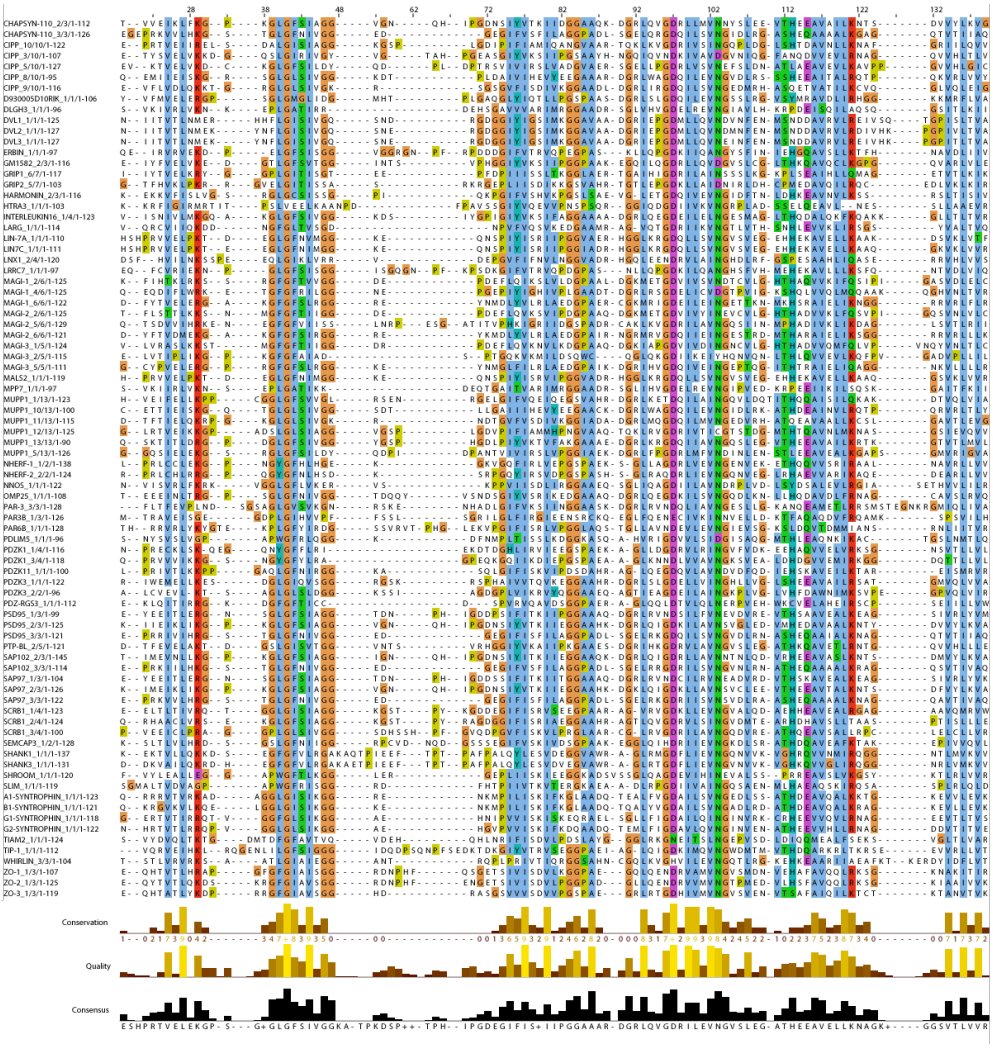
当需要对模型的两分类（即预测PDZ结构域-短肽对是否发生相互作用）能力进行评估时，本章采用各种不同的bootstrap策略（每次随机抽取部分PDZ结构域或者短肽序列）进行测试，并用AUC值进行度量。

4.2.5 PDZ结构域结合短肽的profile的相似性度量

PDZ结构域的相似性可以从PDZ结构的序列角度出发，也可以根据它们各自结合的短肽及其相应的亲和力大小（profile）进行度量。本章针对PDZ结构域的结合短肽profile进行相似性度量时，采用量化的Czekanowski’s指标[157]：



（A）基于a1-syn蛋白的PDZ结构域之38个物理接触对. 图摘自文[95].



（B）PDZ结构域多序列比对结果

图4-4 PDZ结构域-短肽对之特征编码

$$similarity(x_i,x_j)=\frac{2\sum_{k=1}^{n}\min(x_{ik},x_{jk})}{\sum_{k=1}^{n}(x_{ik}+x_{jk})}, \quad (4\text{-}10)$$

其中x_i, x_j为n维向量，代表PDZ结构域，它们中的每一个元素代表了短肽与对应PDZ结构域的相互作用亲和力大小。也就是说，针对发生相互作用的PDZ结构域-短肽对，该元素值为$K_a = 1/K_d$，而对未发生相互作用的PDZ结构域-短肽对，该元素对应为0。该相似性指标度量了与任意两个PDZ结构域同时发生相互作用的短肽之间的亲和力大小的相近程度。该相似性度量指标是目前比较流行的度量定性向量之间的相似性的Dice系数的一个定量化改进。

4.2.6 实验方法

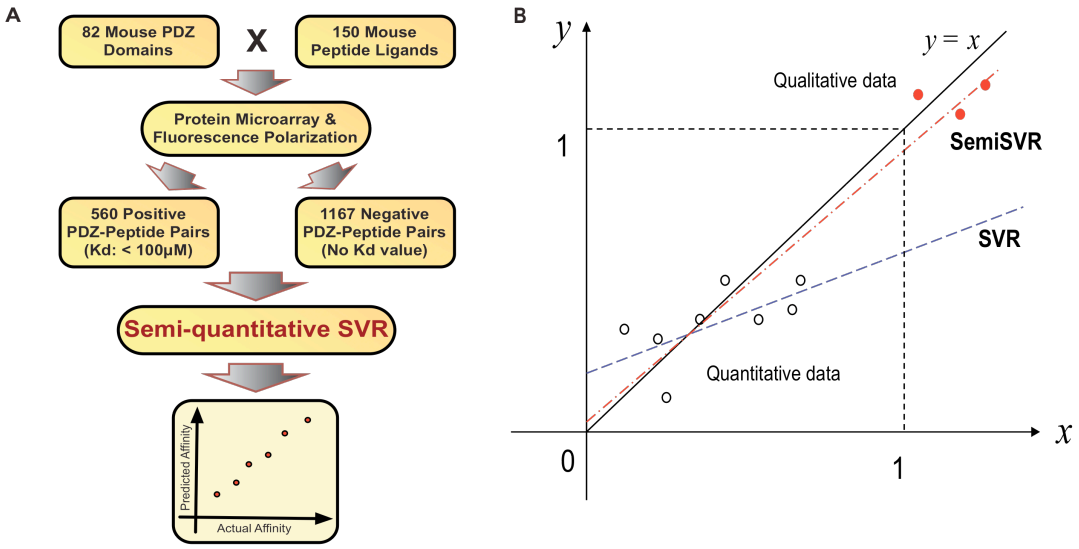
本章为了验证提出的半量化支持向量回归机模型的有效性，还将其对新测得的人类PDZ结构域-短肽相互作用对数据进行预测。本章采用类似于Stiffler等人的方法，通过荧光偏振技术（fluorescence polarization，FP）结合装置实现对人类Scribble蛋白中的第3个PDZ结构域的结构域-短肽相互作用对的亲和力的测定。简言之，首先将该PDZ结构域利用pETM30 plasmid作为GST融合蛋白质进行表达，之后再在Purification buffer (PB; 50 mM Bicine pH 8.25, 200 mM NaCl, 0.5 mM EDTA)中通过GSH亲和力chromatography纯化（purified）并在10 mM GSH (Sigma)环境下洗涤。接着，该蛋白质在S75 gel filtration柱子上利用AKTA FPLC进行进一步的纯化并利用Centricon浓缩机（Millipore）进行浓缩至60-100 mM。然后，将该蛋白质再进行稀释，每次稀释25 ml溶液并均分至384个黑色荧光盘中（Corning）（black fluorescent plate）。之后再将带有荧光标记的短肽蛋白质放至每一个荧光盘中（每一个盘中的荧光强度会对应一种短肽序列）并利用Envision Multi-Label Plate reader (Perkin-Elmer)测得每一个盘中对应的荧光偏振信号。最后，利用Graph-Pad Prism (v3.0)软件中的非线性回归函数计算出PDZ结构域-短肽相互作用之结合亲和力大小。

4.3 实验结果分析

本章工作的目标是通过建立机器学习模型从序列信息角度出发实现对任意给定一个PDZ结构域和短肽序列之间相互作用亲和力的大小的预测。为此，本章建立了一种类似支持向量回归机的机器学习模型进行预测。该模型是建立在最新得到的关于小鼠PDZ结构域-短肽相互作用的高通量数据之上的。

本章获取的数据集既包括正类相互作用对，又包括负类相互作用对。对于回归问题，经典的方法如支持向量回归机，通常仅考虑正类相互作用对（也即需要知道连续的回归值），并不能利用定性的负类相互作用信息。这里假设负类相互作用对可以提供额外的信息进而提高回归模型的性能。为此，我们改进了经典的支持向量回归机模型，提出了一种新的模型——半量化支持向量回归机，该新模型可以同时考虑正类相互作用对和负类相互作用对。整个预测模型框架和算法图示如图4-5。下面将介绍利用该新模型得到的预测结果。如前节所述，基于PDZ结构域和短肽配体的序列信息进行特征编码，并结合它们之间的相互作用亲和力（K_d值），共同作为模型的输入信

息进行建模。本章采用基于PDZ结构域的留一法评估各种回归模型的预测性能。



(A)PDZ结构域-短肽相互作用亲和力预测框架图 (B)半量化支持向量回归机图示

图4-5 PDZ结构域亲和力预测模型框架图

4.3.1 不同特征编码方式的结果比较

本章首先采用前节提及的各种不同的特征编码方式以及各种不同的核函数，利用基于PDZ结构域的留一法进行评估各种不同特征编码方式对应的半量化支持向量回归机的预测性能。此处利用Pearson相关系数和Spearman相关系数作为性能的评价指标。具体的编码方式主要有基于结构的38个氨基酸物理接触对编码（对应为400维的正交向量）；针对PDZ结构域全序列（118AAs）的Profeat特征编码和针对16个结合位点的sparse20编码等，加之针对短肽序列的sparse20编码、zscale编码、5factor和11factor编码等。具体的核函数则包括了针对不同编码方式的线性核函数、高斯核函数和“成对”多项式核函数。其中部分实验用到的编码方式和相应核函数的预测性能见表4-1（在该表中，仅报道了基于Spearman相关系数的结果，关于Pearson相关系数的结果见附录A）。经过比较该表，发现基于“成对”多项式核函数的编码方式具有最佳的预测性能。此外，对于经典的支持向量回归机也具有类似的结果。所以本章最终采用基于“成对”多项式核函数的编码方式作为最终模型用到的编码方式和核函数。事实上，本章采用的“成对”多项式核函数在某种意义上模拟了PDZ结构域内部不同位置的氨基酸对和短肽序列内部不同位置的氨基酸对同时出现的不同倾向。

4.3.2 和经典支持向量回归机的预测性能比较

为了考察负类相互作用对数据对预测PDZ结构域-短肽相互作用的强弱是否起作用，我们首先从“单结构域”模型出发，比较了经典的基于“单结构域”数据的PWM模型、基于“单结构域”

数据的传统的支持向量回归机模型和基于“单结构域”数据的半量化支持向量回归机模型的预测性能。本节考察这些模型预测给定任意一对PDZ结构域-短肽对的相互作用强弱（例如任意两个正类相互作用对或者一个正类相互作用对和一个负类不相互作用对）的能力。为此，测试了关于23个PDZ结构域的全部所有可能的PDZ结构域-短肽对。针对每一个PDZ结构域，每次测试时选取一对PDZ结构域-短肽对，利用余下的PDZ结构域-短肽相互作用对数据构建模型（PWM模型和传统支持向量回归机模型仅采用正类相互作用对数据，而半量化支持向量回归机模型则采用了全部余下的数据）。本节采用预测正确率来衡量模型的性能。具体结果见表4-2。表中结果表明SemiSVR的性能最优（21/23个取到最优结果，平均正确率为0.79比0.72，P值为0.0023），与SVR模型相比，优势更明显。这表明了整合负类相互作用对数据能提高预测性能。

经典的支持向量回归机仅利用正类PDZ结构域–短肽相互作用对数据集，而半量化支持向量回归机则同时整合了正类相互作用对和负类相互作用对数据集。为了比较两者之间的性能，均采用基于“成对”多项式核函数的特征编码和核函数方式。从表4-3的结果可以进一步看出，本章提出的半量化支持向量回归机模型具有比经典支持向量回归机模型更为优越的预测性能，进一步验证了本章所提新模型的有效性。

表4-1 基于不同特征编码的半量化支持向量回归机模型之性能比较

Spearman	38 pairs	BSs-16AA (pairwise)		WholePDZ-118AA (pairwise)		Profeat+11f	Profeat+5f	Profeat+zscale	BS16AA +10AA	BS16AA + 5AA
PDZ domain name	Linear	Linear	Polynomial	Linear	**Polynomial**	RBF	RBF	RBF	Sparse20 RBF	Sparse20 RBF
CHAPSYN-110_2/3	0.946	0.977	0.940	0.922	0.936	0.775	0.806	0.831	0.940	0.934
CHAPSYN-110_3/3	0.602	0.629	0.909	0.715	0.891	0.724	0.735	0.624	0.909	0.870
GM1582_2/3	0.409	0.594	0.488	0.732	0.653	0.318	0.465	0.279	0.488	0.456
HTRA3_1/1	0.236	0.200	0.794	0.261	0.527	0.648	0.661	0.382	0.794	0.127
LIN7C_1/1	0.467	0.394	0.394	0.479	0.612	0.576	0.382	0.406	0.394	0.333
MAGI-2_2/6	0.631	0.707	0.707	0.782	0.700	0.414	0.568	0.475	0.707	0.870
MAGI-2_6/6	0.630	0.560	0.423	0.549	0.637	0.604	0.665	0.368	0.423	-0.129
MAGI-3_1/5	0.730	0.513	0.812	0.730	0.816	0.672	0.478	0.414	0.812	0.676
MALS2_1/1	0.326	0.734	0.538	0.643	0.545	0.434	0.643	0.406	0.538	0.564
OMP25_1/1	0.509	0.265	0.592	0.344	0.528	0.606	0.453	0.389	0.592	0.633
PDZK3_1/1	-0.132	-0.226	-0.244	-0.206	-0.197	-0.597	-0.279	-0.547	-0.244	-0.518
PDZ-RGS3_1/1	-0.002	0.297	0.244	0.323	0.310	0.402	0.481	0.464	0.244	0.099
PSD95_2/3	0.821	0.937	0.958	0.909	0.965	0.804	0.937	0.895	0.958	0.905
PSD95_3/3	0.597	0.599	0.588	0.643	0.747	0.692	0.758	0.896	0.588	0.823
PTP-BL_2/5	0.340	0.500	0.582	0.515	0.356	-0.068	0.379	-0.015	0.582	0.632
SAP102_2/3	0.911	0.799	0.955	0.935	0.968	0.753	0.664	0.740	0.955	0.945
SAP97_1/3	0.455	0.418	0.394	0.394	0.345	0.139	0.382	0.176	0.394	0.406
SAP97_2/3	0.911	0.987	0.934	0.943	0.952	0.938	0.859	0.864	0.934	0.942
SCRB1_3/4	0.370	0.527	0.564	0.479	0.479	0.442	0.491	0.503	0.564	0.515
SHANK1_1/1	0.505	0.709	0.964	0.976	0.976	0.830	0.636	0.539	0.964	0.942
SHANK3_1/1	0.942	0.612	0.358	0.455	0.358	0.285	0.442	0.503	0.358	0.432
G1-SYNTROPHIN_1/1	0.205	0.267	0.033	0.235	0.172	0.346	0.274	0.582	0.033	0.245
ZO-1_1/3	0.606	-0.035	0.727	0.259	0.643	0.657	-0.217	0.301	0.727	0.228
Average Performance	0.522	0.520	0.594	0.566	**0.605**	0.495	0.507	0.455	0.594	0.519

表4-2 基于“单结构域”的SemiSVR、SVR和PWM模型之预测成对PDZ结构域-短肽对之相互作用强弱的性能比较

试验在23个PDZ结构域上进行比较。表中数字代表了平均的预测正确率，加粗的数字表明是最好结果。

PDZ domain	SemiSVR	SVR	PWM
CHAPSYN-110_2/3	**0.75**	0.57	0.71
CHAPSYN-110_3/3	**0.86**	0.60	0.79
GM1582_2/3	**0.74**	0.64	0.68
HTRA3_1/1	**0.73**	0.66	0.70
LIN7C_1/1	**0.89**	0.59	0.76
MAGI-2_2/6	**0.85**	0.55	0.73
MAGI-2_6/6	**0.71**	0.67	0.69
MAGI-3_1/5	**0.71**	0.49	0.64
MALS2_1/1	0.55	0.40	**0.60**
OMP25_1/1	**0.77**	0.63	0.65
PDZK3_1/1	**0.78**	0.64	0.70
PDZ-RGS3_1/1	**0.82**	0.80	0.68
PSD95_2/3	**0.69**	0.37	0.65
PSD95_3/3	**0.82**	0.70	0.80
PTP-BL_2/5	**0.83**	0.60	0.77
SAP102_2/3	**0.81**	0.63	0.66
SAP97_1/3	**0.74**	0.57	0.69
SAP97_2/3	**0.74**	0.50	0.71
SCRB1_3/4	**0.84**	0.59	0.75
SHANK1_1/1	**0.91**	0.88	0.81
SHANK3_1/1	**0.88**	0.82	0.80
G1-SYNTROPHIN_1/1	**0.87**	0.58	0.79
ZO-1_1/3	**0.75**	0.51	**0.75**
Average Performance	**0.79**	0.61	0.72

表4-3 基于“多结构域”的SemiSVR和SVR模型的预测性能比较

该比较是采用基于PDZ结构域的留一法对23个PDZ结构域进行测试。两个模型均采用最佳的特征编码和模型参数（即基于118AAs的PDZ结构域和10AA的短肽序列，采用阶数为2的“成对”多项式核函数）。采用Spearman和Pearson相关系数度量结果性能。加粗的数字表明是最好结果。

Performance Measure	SemiSVR	SVR
Spearman	**0.605**	0.501
Pearson	**0.653**	0.574

4.3.3 与现有方法比较

目前针对PDZ结构域-短肽相互作用亲和力预测的工作仅限于Chen等人于2008年发表的基于Bayesian思想的Additive模型。此处将与该算法进行比较，具体结果见表4-4。从该表可以看出，基于“成对”多项式核的半量化支持向量回归机具有最好的预测性能（无论是从Spearman相关系数还是Pearson相关系数）。由于Chen等人采用的是基于38个相互作用氨基酸对进行编码PDZ结构域-短肽对，为了公平的比较，本章也采用相同的特征编码方式进行编码。在此情形下，本章提出的半量化支持向量机仍有更好的预测性能，这表明本章提出的算法模型更为有效的整合了正类样本信息核负类样本信息，比Chen的方法更具有优势。

表4-4 SemiSVR模型与已有方法之结果比较

该比较是采用基于PDZ结构域的留一法对23个PDZ结构域进行测试。采用Spearman和Pearson相关系数度量结果性能.表中第2-4列代表了基于118AAs的特征编码、基于38个物理接触对的特征编码的SemiSVR模型和Chen等人提出的Additive模型。加粗的数字表明是最好结果。在基于38个物理接触对的特征编码的SemiSVR模型中选取了线性核函数。Chen等人的模型采用文[95]中相应程序重新运算获得。

Performance measure	**Spearman correlation/Pearson correlation**		
PDZ domain	SemiSVR WholePDZ-118AA	SemiSVR 38pairs	Chen
CHAPSYN-110_2/3	0.94/**0.94**	**0.95**/0.93	0.80/0.79
CHAPSYN-110_3/3	**0.89/0.88**	0.60/0.57	0.59/0.50
GM1582_2/3	**0.65/0.58**	0.41/0.35	0.36/0.19
HTRA3_1/1	**0.53/0.65**	0.24/0.36	0.20/0.13
LIN7C_1/1	**0.61/0.68**	0.47/0.56	-0.37/-0.17
MAGI-2_2/6	**0.70**/0.77	0.63/**0.78**	0.11/0.21
MAGI-2_6/6	**0.64/0.69**	0.63/0.52	0.28/0.17
MAGI-3_1/5	**0.82/0.88**	0.73/0.68	0.54/0.52
MALS2_1/1	**0.55/0.61**	0.33/0.37	0.17/0.15
OMP25_1/1	**0.53**/0.50	0.51/**0.51**	0.32/0.37
PDZK3_1/1	-0.20/**0.04**	**-0.13**/0.02	-0.22/0.02
PDZ-RGS3_1/1	**0.31**/0.03	-0.002/-0.05	-0.08/**0.07**
PSD95_2/3	**0.97/0.92**	0.82/0.87	0.53/0.66
PSD95_3/3	**0.75/0.88**	0.597/0.68	0.22/0.17
PTP-BL_2/5	**0.36**/0.40	0.34/**0.53**	0.18/0.16
SAP102_2/3	**0.97/0.94**	0.91/0.92	0.91/0.94
SAP97_1/3	0.34/**0.76**	**0.46**/0.63	-0.16/0.14
SAP97_2/3	**0.95/0.95**	0.91/0.92	0.77/0.85
SCRB1_3/4	0.48/0.69	0.37/0.47	**0.697/0.78**
SHANK1_1/1	**0.98/0.98**	0.51/0.44	0.95/0.96
SHANK3_1/1	0.36/0.51	**0.94/0.91**	0.69/0.70
G1-SYNTROPHIN_1/1	0.17/0.13	0.21/0.16	**0.52/0.48**
ZO-1_1/3	**0.64/0.65**	0.61/0.64	0.26/0.16
Average Performance	**0.61/0.65**	0.52/0.56	0.36/0.39

4.3.4 结构域的序列相似性影响预测性能

由表4-4可以看出，半量化支持向量回归机在有些PDZ结构域上有着非常卓越的预测性能，但也出现了对个别几个PDZ结构域几乎无法预测的情形。为了究其原因，首先考察到底是PDZ结构域序列的哪些区域影响着最终模型的预测性能。为此，除了之前采用的全序列（118AAs）和由16个结合位点构成的伪序列之外，还增加了采用由10个核心结合位点[92]构成的伪序列来代表

PDZ结构域，并构建半量化支持向量回归机模型。通过最终的性能比较，发现采用PDZ结构域的全序列（118AAs）的半量化支持向量机模型得到了最佳的预测性能。诚然，采用基于由16个结合位点或者10个核心结合位点构成的伪序列的半量化支持向量机模型也得到了非常接近的预测性能，具体结果见表4-5。进一步，为了考察是否由于随机因素引起的模型性能差异，我们从非结合位点中随机（20次）选取了16个或者10个氨基酸长度的伪序列（不考虑多序列比对后超过75%“插入”的位置）作为PDZ结构域的代表序列，并基于此伪序列进行特征编码进而训练半量化支持向量回归机。与基于16个结合位点或者10个核心结合位点的半量化支持向量回归机模型相比，基于随机选择的非结合位点的模型的预测性能要显著地差（性能相差6%左右，具体结果参加表4-6）。综合上述分析，表明了由结合位点构成的伪序列提供了最主要的信息，同时其他非结合位点也提供了额外的信息。

表4-5 基于PDZ结构域的不同子序列对应的SemiSVR预测性能比较

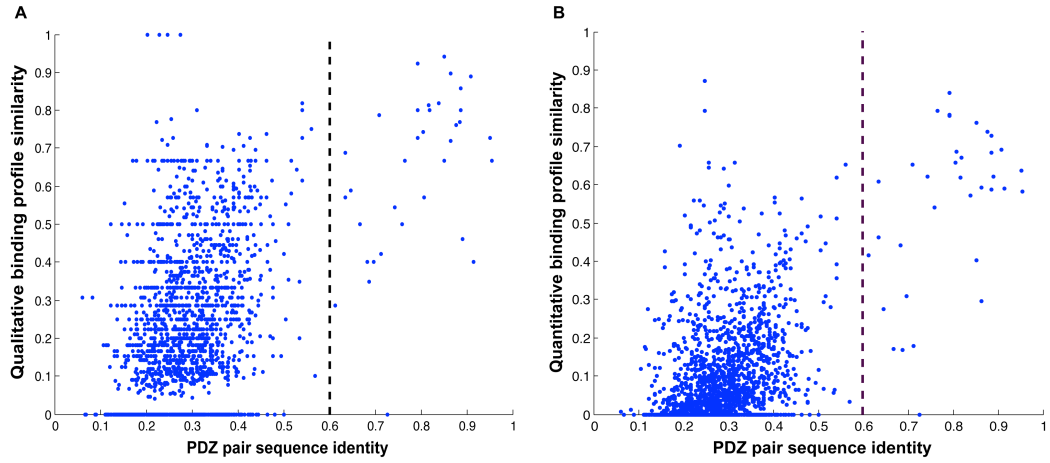
性能指标	特征编码	SemiSVR性能
Spearman	WholePDZ-118AA	**0.605**
	BindingSite-16AA	0.594
	CoreBindingSite-10AA	0.594
Pearson	WholePDZ-118AA	**0.653**
	BindingSite-16AA	0.636
	CoreBindingSite-10AA	0.649

表4-6 基于PDZ结构域的结合位点与随机选择同长度位点的SemiSVR预测性能比较

Random10BS	Average performance on 23 PDZ domains		Random16BS	Average performance on 23 PDZ domains	
Random No.	Spearman	Pearson	Random No.	Spearman	Pearson
1	0.492	0.543	1	0.569	0.612
2	0.583	0.620	2	0.571	0.590
3	0.553	0.605	3	0.598	0.621
4	0.590	0.634	4	0.583	0.578
5	0.521	0.566	5	0.577	0.600
6	0.554	0.596	6	0.561	0.601
7	0.535	0.580	7	0.595	0.630
8	0.520	0.589	8	0.578	0.603
9	0.525	0.589	9	0.591	0.621
10	0.576	0.620	10	0.514	0.558
11	0.585	0.608	11	0.561	0.616
12	0.589	0.656	12	0.533	0.598
13	0.582	0.617	13	0.553	0.587
14	0.586	0.624	14	0.566	0.600
15	0.551	0.584	15	0.569	0.600
16	0.532	0.579	16	0.529	0.573
17	0.569	0.619	17	0.554	0.598

18	0.572	0.624	18	0.554	0.610
19	0.555	0.603	19	0.523	0.561
20	0.526	0.590	20	0.538	0.585
Average	0.555	0.602	Average	0.561	0.597
Std	0.029	0.026	Std	0.024	0.019
T-test	Difference	Difference	Difference	Difference	Difference
10corebinding sites	**0.594**	**0.649**	16binding sites	**0.594**	**0.636**
P-value	8.02E-06	1.27E-05	P-value	6.69E-06	2.89E-08

由上面的分析，可以大胆推测基于序列编码的半量化支持向量回归机模型的预测性能在一定程度上依赖于PDZ结构域之间的序列相似性。为此，首先考察PDZ结构域之间的序列相似性与PDZ结构域的结合profile相似度之间的关系。分别采用针对定性化数据的dice度量和针对定量化的Czekanowski度量来度量任意两个PDZ结构域之间的结合profile相似度。序列相似度与结合profile相似度之间的关系图见图4-6。由图可以看出，当PDZ结构域之间的序列相似度高于60%，随着序列相似度的增加，其结合profile的相似度也随之增加。而这个结果也与Tonikian等人的发现一致。



（A）定性化结合profile相似度 （B）定量化结合profile相似度

图4-6 PDZ结构域序列相似度与结合profile相似度的散点图

此外，我们考察了各个PDZ结构域的预测性能和与之近邻的PDZ之间的序列相似性之间的关系。这里采用118个氨基酸长度序列作为PDZ结构域的特征序列进行度量序列之间的“一致性”（Identity）。研究发现这两者之间有着显著的正相关性（Pearson相关系数为0.498，P-value为0.0157），见图4-7。当采用16个结合位点或者10个核心结合位点来度量PDZ结构域之间的相似性时，也得到了类似的结果。为了更为具体地考察“近邻”序列相似性对最终模型的性能影响，我们采用不同的序列相似度阈值进行去“冗余”序列的模拟，也即对每个待预测的PDZ结构域（及其相关的短肽相互作用对），首先选择去掉与该序列具有序列相似度高于给定阈值的所有序列（及其相关的短肽相互作用对），再基于剩余的PDZ结构域-短肽相互作用对数据构建半量化支持向量回归机模型。通过该试验模拟，发现半量化支持向量回归机的预测性能随着“近邻”序列的阈值提高而有所降低，具体结果参见图4-8。上述分析表明了现有的半量化支持向量回归机模型的预测性能将在

很大程度上依赖于测试PDZ结构域是否在训练集中有“充分近邻”的PDZ结构域。

4.3.5 基于结构域家族的模型具有较好的预测性能

本章得到的最终模型是基于整个PDZ结构域家族的，也就是说整合了全部的PDZ结构域以及短肽相互作用对数据构建了一个统计的模型。但是如前面所述，通常有两种策略构建预测模型，基于多个结构域的是一类（“全局模型”），基于单个结构域的是另一类（“局部模型”）。而且前节的结果也表明了本章提出的半量化支持向量回归机模型在一定程度上依赖于测试PDZ结构域是否在训练集中有“充分近邻”的PDZ结构域。那么本章构建的模型是否比基于单个近邻结构域的模型的预测性能更好呢？为此，本章针对基于每一个PDZ结构域的测试集，构建了基于其最近邻PDZ结构域数据的半量化支持向量回归机模型。此外，本章还构建了基于“最近邻”PDZ结构域的PWM模型。此处采用了“Identity”和“Blosum62”两种标准度量PDZ结构域之间的相似度，并且选取了全序列和16个结合位点、10个核心结合位点构成的伪氨基酸序列作为PDZ结构域的序列，并据此确定“最近邻”PDZ结构域。各种不同模型的预测性能具体参见表4-7。从表中可以看出基于“最近邻”的半量化支持向量回归模型优于其他各种基于PWM的模型，这可能是因为半量化支持向量回归模型整合负类相互作用信息。此外，从表中也发现基于整个PDZ结构域家族的模型优于其他所有的模型，这表明了虽然“最近邻”PDZ结构域信息对于最终的基于“多结构域”半量化支持向量回归机模型起到了重要的作用，但是整合其他非近邻的PDZ结构域信息对最终的模型也有一定的帮助。

表4-7 SemiSVR与其他基于局部信息的模型之间的预测性能比较

其中局部信息模型主要含有基于最近邻的SemiSVR模型，有基于最近邻的PWM模型。在度量时采用了不同的PDZ子序列和不同的序列相似性度量标准（序列一致性和Blosum62评分）。

预测模型		Spearman	Pearson
SemiSVR	118AA	0.605	0.653
Nearest Neighbor SemiSVR	118AA	0.471	0.487
Naïve PWM transfer (Identity)	118AA	0.303	0.323
	16BSs	0.305	0.319
	10BS	0.326	0.303
Naïve PWM transfer (Blosum62)	118AA	0.305	0.311
	16BSs	0.296	0.274
	10BS	0.354	0.286

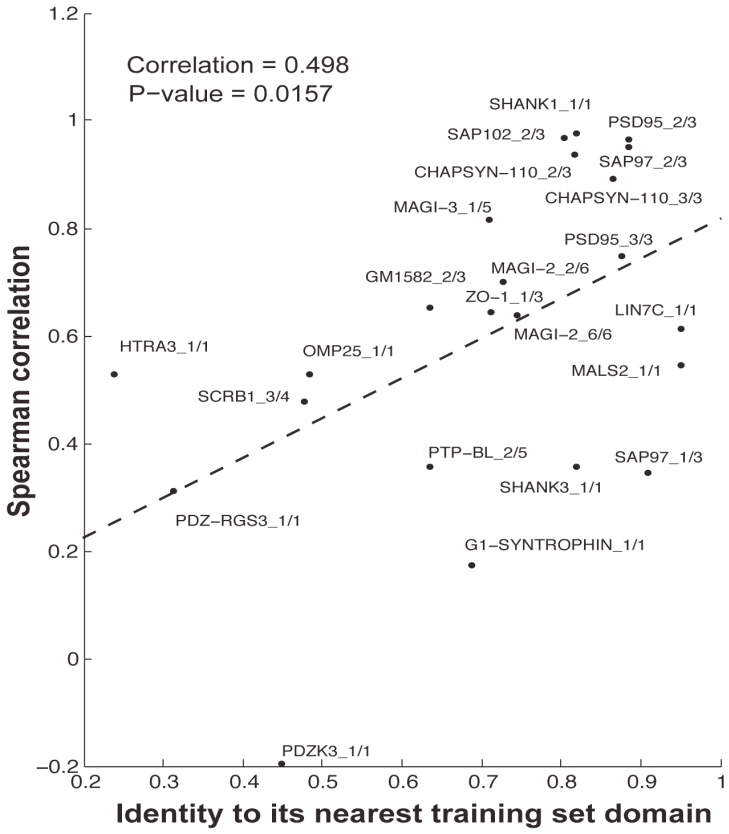


图4-7 预测性能与到最近邻PDZ结构域的序列相似度之间的散点图

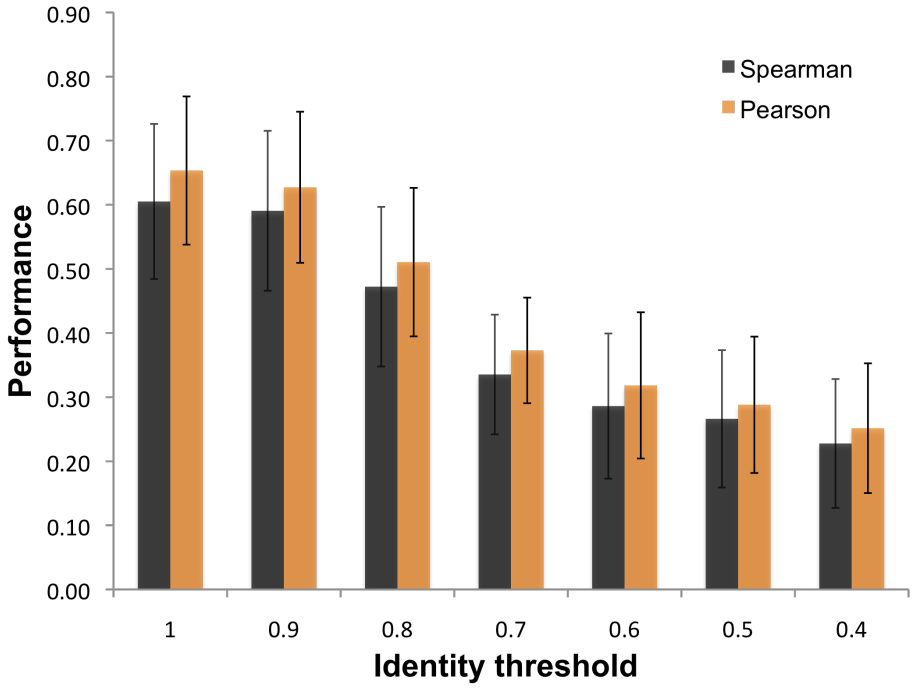


图4-8 预测性能随序列相似度阈值的减小而降低

4.3.6 独立测试集（人类PDZ结构域）

前面已经从交叉验证的角度给出了本章提出的半量化支持向量回归机模型的优势。接下来，还将测试新模型在新近测得的未用于训练的PDZ结构域-短肽相互作用对数据上的预测结果。测试的数据集是关于Scribble蛋白中的第3个PDZ结构域与人类中57个蛋白质短肽配体之间的相互作用数据，具体实验数据见表4-8。该实验测得了36个正类相互作用对（K_d值低于100*u*M）。测试结果表明新模型在该正类数据集上的预测性能较佳，Spearman相关系数高达0.74（P值为8.85e-7）。此外，我们在线虫和果蝇PDZ结构域-短肽相互作用数据上测试也得到了类似的结果。这类数据集是Chen等人在文[95]中测得的，分别在果蝇和线虫中测定了含有7个PDZ结构域的结构域-短肽相互作用对数据。

表4-8 人类Scribble蛋白之PDZ结构域-短肽相互作用数据及预测结果

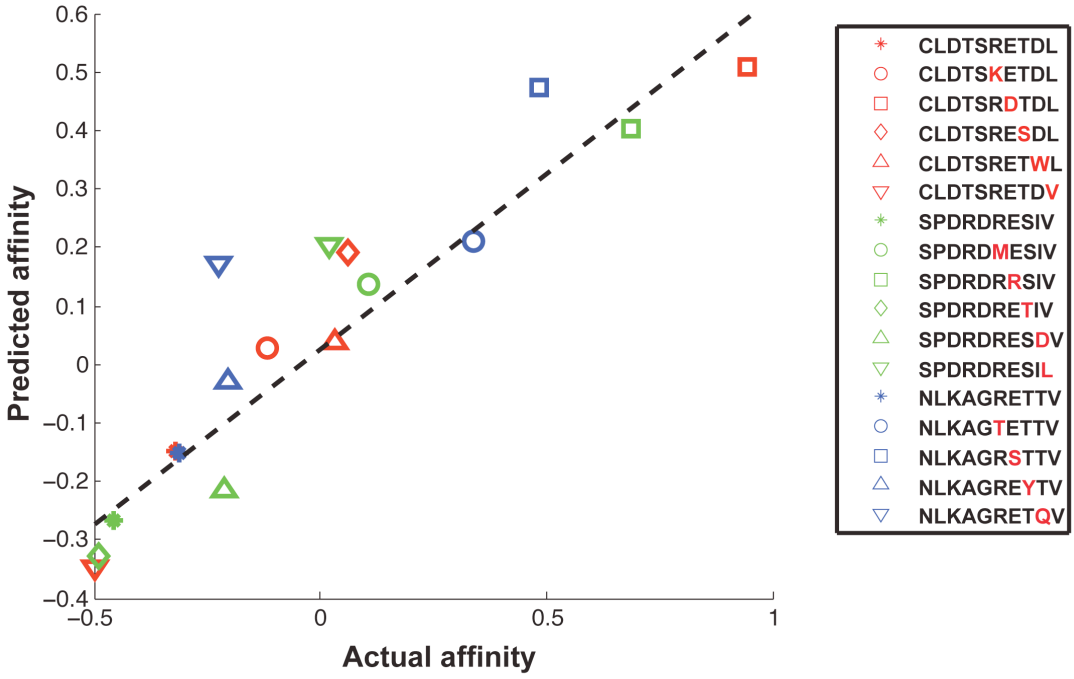
蛋白质名称	蛋白质ID	C-末端序列	实际K_d值（*u*M）	标准差	预测的SemiSVR值
ARD50	BC024725.1	SFNYKKETPL	1.5	0.1	0.543
MAPK12	PV3654	GARVSKETPL	1.7	0.1	-0.052
ARHGEF16	NM_014448.2	MERLRVETDV	2.1	0.1	0.615
MCC	NM_002387.1	SRPHTNETSL	2.4	0.1	0.466
ABR	NM_001092.3	RNTLYFSTDV	2.9	0.1	0.899
STK29	BC024291.1	KVATSYESSL	3.2	0.2	0.705
FAM105B	NM_138348.3	PVRVCEETSL	3.4	0.2	0.446
b-PIX	NP_003890	NDPAWDETNL	4.9	0.4	0.628
APC	NP_000029	HSGSYLVTSV	6.7	0.4	0.521
VANGL2	NP_065068	VMRLQSETSV	8.8	0.5	0.553
PRKCA	P2227	FVHPILQSAV	9.8	0.5	0.710
AHDC1	BC002677.1	PEDTFTVTSL	10.4	0.6	0.814
LPP	NP_005569	VLTAKASTDL	11.3	0.5	0.960
FOXl1	NM_144769.1	VLYPREGTEV	11.7	0.8	0.930
KCNJ10	BC034036.1	SALSVRISNV	13.2	1	0.910
TANK	NM_133484.1	VDIASAESSI	13.6	0.6	0.920
ZO2	NP_004808	QSARYRDTEL	13.7	0.6	0.869
ZNF654	NM_018293.1	SSAQPSETIL	18.3	1	0.523
KIRREL2	BC007312.1	PSHPRLQTHV	21.3	2.4	0.650
MCM7	BC009398.1	NASRTRITFV	24.1	2.4	0.867
RPS6KA2	NM_001006932	GMKRLTSTRL	25.2	2.1	0.987
C11orf52	NM_080659.1	RYDSKNGTLV	26.5	2.4	0.952
RPS6KA1	NM_001006665	RVRKLPSTTL	27	3.2	1.061
ANKS4B	NM_145865.1	QPGQLVDTSL	28.2	2	0.926
RASL11B	NM_023940.1	SAKVRTVTSV	30	3.5	0.709
SYNJ2BP	BC007704.1	WAFMRYRQQL	32	1.8	1.443
KCNA6	NM_002235.2	YAEKRMLTEV	34.8	3.6	0.859

EPHA8	PV3844	DPELEALHCL	38.8	6.3	1.324
FAM126B	NM_173822.1	SFNMQLISQV	39.3	5.1	1.103
DSC54	NM_016644.1	ILRKSTTTTV	40.1	5.2	0.848
EPHA5	PV3840	VQLVNGMVPL	45.6	8.7	1.390
SRC	NM_005417.3	EPQYQPGENL	48.6	4.4	1.400
PDGFRA	NM_006206.3	SSDLVEDSFL	52	7.1	1.129
STK16	BC053998.1	PAPGQHTTQI	52.2	5.6	1.191
TPM2	NM_003289.3	DNALNDITSL	54.4	11.1	0.881
GLO1	BC001741.1	LNPNKMATLM	59.5	6.6	1.304
ZADH2	NM_175907.3	ELPHSVNSKL	100	0	1.108
PDGFRB	NM_002609	PRAEAEDSFL	100	0	1.037
MPG	BC014991.1	DRVAEQDTQA	100	0	1.335
PRKCB1	P2281	YTNPEFVINV	100	0	1.252
TBK1	PV3504	DGGLRNVDCL	100	0	1.356
MTERFD1	NM_015942.3	QDFEKFLKTL	100	0	1.322
FLT1	NM_002019.1	NSVVLYSTPPI	100	0	1.498
PSMA8	BC042820.1	AEKKKSKKSV	100	0	1.237
EPHA7	PV3689	LHLHGTGIQV	100	0	1.007
EPHA2	PV3688	DQVNTVGIPI	100	0	1.544
DIRAS1	NM_145173.1	DRVKGKCTLM	100	0	1.213
BEGAIN	NM_020836.2	KAQLYGTLLN	100	0	1.540
MUSK	PV3834	CERAEGTVSV	100	0	1.269
EPHA3	PV3359	TQSKNGPVPV	100	0	1.365
TRIM21	NM_003141.2	NIGSQGSTDY	100	0	1.252
LIMD1	NM_014240.1	SSTALHQHHF	100	0	1.518
C19orf57	BC012945.1	IPRGDPPWREL	100	0	1.298
PTE1	NM_005469.2	VKPQVSESKL	100	0	0.950
UBXD1	NM_025241.1	PELLSAIEKLL	100	0	0.914
ACBD6	NM_032360.1	VLQRHTTGKA	100	0	1.584
PACAP	BC021275.1	SEKVSATREEL	100	0	1.366

4.3.7 预测由短肽序列单点变异引起的亲和力变化

为了进一步测试本章提出的半量化支持向量回归机的性能，我们将对短肽序列进行单点变异并预测变异之后的新短肽序列与PDZ结构域之间的亲和力的大小。采用与a1-synPDZ结构域具有相互作用的3个野生型短肽序列（Kv1.5、Nav1.5和KIF1B）作为测试对照集，并分别对上述3个短肽序列进行5次单点变异，共测得15对PDZ结构域-短肽相互作用对，并且其中关于KIF1B短肽的一个单点变异之后得到的短肽已不再与a1-synPDZ具有相互作用。该数据集最初也是由Chen等人利用荧光偏振技术测得。采用前面训练得到的半量化支持向量回归机模型对这15对PDZ结构域-短肽序列对进行亲和力的预测，发现该模型能成功地预测由于单点变异导致的亲和力大小的变化，也即可以准确地预测变异之后的短肽与a1-synPDZ结构域相互作用亲和力相对于野生型而言是变

强还是变弱。如图4-9所示，模型预测得到的亲和力与实验得到的亲和力之间具有相当高的相关系数（Spearman相关系数为0.921，P值为1e-16；Pearson相关系数为0.922，P值为1.414e-07）。这表明本章提出的模型能正确地预测由短肽序列进行单点变异引起的亲和力大小的变化。



野生型的短肽采用“＊”号表示，其他标记为红色的氨基酸为相应位置进行单点变异后的氨基酸。

图4-9 短肽单点变异后SemiSVR的预测性能图示

4.3.8 不同的确定正类相互作用对的阈值

本章用到的数据来自Stiffler等人的工作，在他们的实验中，采用了100*u*M作为正类和负类相互作用对区分的阈值。本章的新模型表明在该阈值条件下，模型具有较好的预测性能。为了考察新模型是否具有普适性，我们人为地改变不同的阈值来定义PDZ结构域-短肽相互作用的正类集合，例如取定50*u*M、20*u*M和10*u*M等。通过实验表明，改变不同的阈值定义正类和负类相互作用对，本章提出的新模型都能得到较好的预测性能。由于改变不同的阈值，对部分PDZ结构域而言，其相互作用的短肽个数会有所减少，有的甚至会少于10个短肽。与前面类似，对于少于10个相互作用短肽的PDZ结构域，暂不统计其预测性能。由于大部分正类PDZ结构域-短肽相互作用对的K_d值均在50*u*M之内，所以取定50*u*M对应的新模型的预测结果与取定100*u*M基本相同，具体的结果此处不再罗列。对于取定20*u*M和10*u*M的结果，具体可以参见表4-9 (A)和(B)。

表4-9 基于确定PDZ结构域-短肽相互作用对的不同阈值的SemiSVR预测性能

（A）基于确定正负类阈值为20*u*M的SemiSVR模型的预测性能

PDZname	#Peptides	Spearman	Pearson
CHAPSYN-110_2/3	15	0.932	0.931

CHAPSYN-110_3/3	14	0.516	0.837
MAGI-2_2/6	13	0.654	0.791
MAGI-3_1/5	17	0.775	0.899
OMP25_1/1	30	0.501	0.531
PDZK3_1/1	11	0.191	0.447
PDZ-RGS3_1/1	12	0.294	-0.172
PSD95_2/3	11	0.973	0.935
PSD95_3/3	10	0.661	0.945
SAP102_2/3	13	0.962	0.914
SAP97_1/3	10	0.200	0.754
SAP97_2/3	10	0.964	0.971
SHANK3_1/1	10	0.491	0.607
G1-SYNTROPHIN_1/1	10	0.042	0.027
ZO-1_1/3	11	0.636	0.638

(B) **基于确定正负类阈值为10*u*M的SemiSVR模型的预测性能**

PDZname	#Peptides	Spearman	Pearson
CHAPSYN-110_2/3	11	0.918	0.902
CHAPSYN-110_3/3	10	0.794	0.746
MAGI-3_1/5	12	0.699	0.868
OMP25_1/1	21	0.443	0.437
SAP102_2/3	11	0.936	0.873

4.3.9 分类问题性能分析

本章的重点是预测PDZ结构域-短肽相互作用对之间的亲和力大小，这是一个极其困难而又有挑战性的工作。另外一个较为容易的工作是仅考虑PDZ结构域-短肽相互作用对之间是否相互作用（两分类问题），也即定性地预测给定一个PDZ结构域-短肽相互作用对，它们之间是否相互作用。我们希望本章提出的新模型在相对简单的分类问题上也能有较好的预测性能。为此，仍旧通过基于PDZ结构域的留一法进行交叉验证，并仅对具有足够多相互作用短肽序列的结构域进行预测（此处仍对应为23个PDZ结构域）。我们采用ROC曲线下方对应的面积AUC值作为评价分类性能的指标。在对上述23个PDZ结构域进行交叉验证之后，模型得到的平均AUC值高达0.88（如图4-10(A)）。这表明新模型也能很好地预测PDZ结构域-短肽相互作用对之间是否发生相互作用。

为了和Chen等人发表的结果进行比较，采用与Chen等人相同的交叉验证方式来验证新模型的分类性能。在Chen等人的工作中，他们采用bootstrap测试法进行测试，主要有3种：1）基于PDZ结构域的bootstrap测试策略，即每次测试时，从82个PDZ结构域种随机选取12％的PDZ结构域（约10个PDZ结构域）及其相关的短肽相互作用对作为测试集，用剩余的PDZ结构域对应的PDZ结构域-短肽相互作用对作为训练集；2）基于短肽的bootstrap测试策略，即每次从所有的短肽中随机选取8％的短肽序列（约12个短肽序列）及其相关的PDZ结构域作为测试集，用剩余的短肽涉及的PDZ结构域-短肽对作为训练集；3）同时基于PDZ结构域和短肽的bootstrap测试策略，也即每次随机取出12％的PDZ结构域和8％的短肽序列，用剩余的PDZ结构域和短肽序列对应的PDZ结构域

-短肽对作为训练集。针对上述3种不同策略得bootstrap测试，本章提出的新模型也得到了较好的结果，与Chen等人的结果相当，对应的AUC值分别为0.862±0.016、0.853±0.021和0.891±0.007，如图4-10(B)所示。

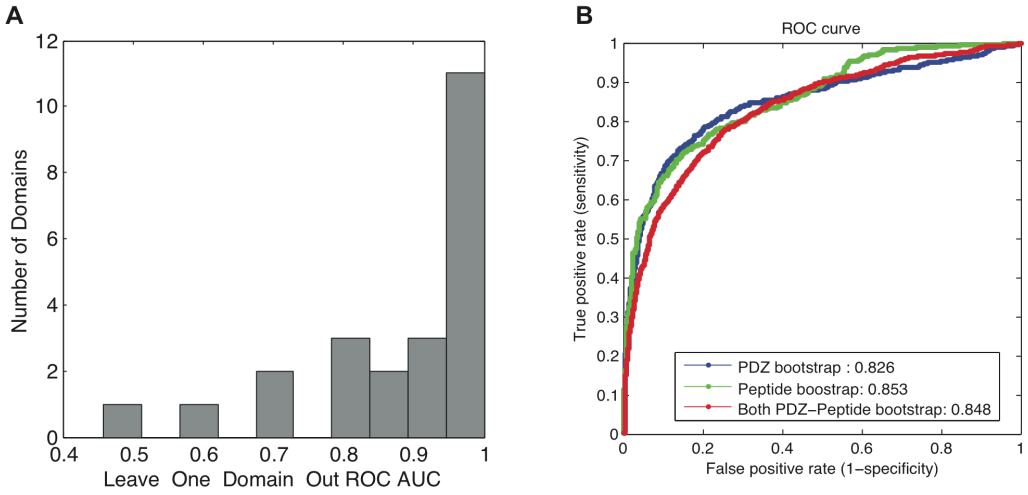


图4-10 SemiSVR的分类预测性能

（A）基于PDZ结构域的留一法预测性能AUC值之分布，（B）基于PDZ结构域、短肽序列、PDZ和短肽序列的Bootstrap测试结果。

本章采用的数据集是基于蛋白质芯片技术结合荧光偏振技术而获得的含有关于PDZ结构域-短肽相互作用对亲和力大小的实验数据。然而，在上述实验中，仅基于蛋白芯片技术还获得了大量的“负样本”PDZ结构域-短肽相互作用对（大约16,607对，这些样本对的亲和力并未获得荧光偏振技术的最后测定），该样本集可以作为本章提出的新模型的一个独立测试集进行测试模型的分类性能。利用由上述训练集训练得到的半量化支持向量回归机模型来预测这些蛋白芯片“负样本”点，结果发现大于92%的上述PDZ结构域-短肽相互作用对被预测为“负类”（亲和力大于100*u*M），仅仅含1,173对PDZ结构域-短肽相互作用对是被预测为“正类”。这个结果表明了新的模型可以较好地预测“负类”样本点，进一步验证了新模型的有效性。

4.3.10 物理化学特性分析

此外，本章还分析了不同的物理化学性质对PDZ结构域-短肽相互作用的不同影响。为此，利用“11-factor”的编码方式对PDZ结构域-短肽相互作用对进行编码，并由此分析这些物理化学特性对预测PDZ结构域-短肽相互作用强弱的影响。根据文献[85]中对“11-factor”的描述，这种“11-factor”编码涉及的物理化学特性主要是：steric parameter, number hydrogen bond donors, hydrophobicity scale, hydrophilicity scale, average accessible surface area, van der Waals parameter R0 (relating to amino acid volume), van der Waals parameter epsilon (relating to number of heavy atoms in a side chain), free energy of solution in water, average side chain orientation angle, polarity和isoelectric point。首先，分别利用这11个物理化学性质对每个氨基酸进行编码，得到了11种描述相同的PDZ结构域-短肽相互作用对的方式。然后针对每一种编码方式得到的数据集，构造半量化支持向量回

归机，并且利用基于PDZ结构域的留一法来评价每一个半量化支持向量回归机的预测性能。通过比较分析，发现对PDZ结构域-短肽相互对的亲和力的预测起到重要影响的物理化学因素由强到弱的顺序依次为：isoelectric point, hydrophilicity scale, polarity, average accessible surface area, van der Waals parameter epsilon和steric parameter。这就表明这些物理化学特性可能是调节PDZ结构域-短肽相互作用强弱的几个物理化学特性，具体结果见图4-11。

另外一个非常有意义的工作是考察上述物理化学特性与PDZ结构域-短肽相互作用对的结构特征之间的关系。为了考虑结构特征，仍利用Chen等人提出的基于三维结构的38个相互作用接触对的特征描述方式。首先将所有PDZ结构域-短肽相互对分成3组：不相互作用对（负类点），弱相互作用对（亲和力K_d值在10*u*M和100*u*M之间的相互作用对）和强相互作用对（亲和力K_d值低于10*u*M的相互作用对）。对任意一对PDZ结构域-短肽相互作用对中（38个接触对中）的任一氨基酸相互接触对，将该接触对用如下的7个值进行表示：1）PDZ结构域上的氨基酸对应的Hydrophilicity值＋短肽上的氨基酸对应的Hydrophilicity值，2）PDZ结构域上的氨基酸对应的平均ASA值＋短肽上的氨基酸对应的平均ASA值，3）PDZ结构域上的氨基酸对应的Polarity值＋短肽上的氨基酸对应的Polarity值，4）PDZ结构域上的氨基酸对应的Steric参数值＋短肽上的氨基酸对应的Steric参数值，5）PDZ结构域上的氨基酸对应的氢键donor值＋短肽上的氨基酸对应的氢键donor值，6）PDZ结构域上的氨基酸对应的范德华参数epsilon值＋短肽上的氨基酸对应的范德华参数epsilon值，以及7）PDZ结构域上的氨基酸对应的Isoelectric point值－短肽上的氨基酸对应的Isoelectric point值。之后，根据划分好的3组PDZ结构域-短肽相互作用对针对上述7个物理化学值在38个氨基酸接触对上进行统计。结果表明，在“不相互作用对”和“弱相互作用对”两组PDZ结构域-短肽对之间有多个接触对上存有显著的差异。这些位置很有可能就是联系物理化学特征和结构特征存有很大关联的位置，共同决定着PDZ结构域-短肽之间的相互作用特性。然而，我们也发现仅有几个位置上在“弱相互作用对”和“强相互作用对”之间存在有差异。具有显著统计差异的这几个位置主要集中在短肽的“0”和“-2”位置，这个结果与现有的对PDZ结构域-短肽相互作用对的结构认识（通常PDZ结构域的结合pocket表面就是与短肽的这两个位置的氨基酸进行结合）一致[138]。

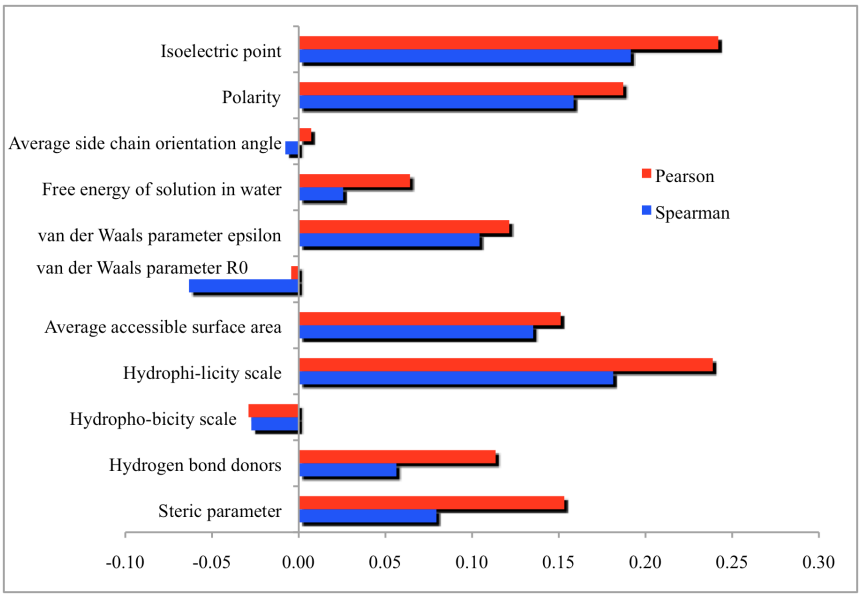


图4-11 基于“11factor”之不同生物化学特性编码的SemiSVR模型性能比较

4.4 结论和工作展望

推断由短肽识别模块介导的蛋白质-短肽相互作用之间的亲和力能有助于人们更深入地理解生物体内细胞活动的过程[141]。然而，到目前为止，这方面的工作开展得还比较少，其中原因之一是缺乏大规模的测得亲和力大小的结构域-短肽相互作用对的数据。本章的工作和结果表明从PDZ结构域的序列信息预测PDZ结构域-短肽相互作用对的相互作用强弱是可行的。同时本章的结果还表明了新模型能有效地整合正类和负类相互作用对数据，并且相比仅利用正类数据的模型，整合负类相互作用数据的模型性能有所改进。此外，本章的新模型也能较为有效地识别正类相互作用对和负类相互作用对。

要精确地预测每一对PDZ结构域-短肽相互作用对的亲和力大小是一件极为困难而富有挑战的工作。本章在这方面进行了初步的尝试并取得了一定的成效，但是距离完全地实现精准的定量化的预测还有很长的一段距离，还是许多问题值得我们进一步的考虑和研究。

通过统计发现，有些PDZ结构域享有一致的伪氨基酸序列，并且可以与相同的短肽配体相互作用，但是它们在相互作用的强弱上却存在很大的差别。例如, PDZ结构域Dvl1 (1/1)和Dvl3 (1/1)在16个结合位点上具有一致的氨基酸，并且均与短肽配体Caspr4相互作用，但是各自的相互作用亲和力则分别为79.298*u*M和30.756*u*M，具有显著差异。这个结果表明在主要的结合位点之外的其他位置氨基酸组成，对PDZ结构域与短肽相互作用也具有一定的影响。在前人的工作中，也已经有这方面的报道，Lockless等人在文[158]中指出部分非结合位点能与结合位点一块起作用影响到PDZ结构域-短肽的相互作用强弱。而且还发现在这些非结合位点发生氨基酸变异的话将会影响到结合位点区域的结构，进而影响PDZ结构域-短肽相互作用强弱。然而，在模型结果分析里，发现采用PDZ结构域全序列编码的模型和采用部分结合位点的伪序列（如由16个结合位点或者10个核心结合位点构成）编码的模型预测性能相当或者稍好一点。这可能是由于本章采用的基于序列的编码方式并未完全有效地获取PDZ结构域-短肽相互作用的规律，需要进一步地发掘表征序列-结构之间关系的有效特征。同时，描述PDZ结构域-短肽相互作用对的其他信息如PDZ结构域的结构特征，协同进化位点特征等也将可能对构建新模型起到关键的作用，此外，若能进一步获取关于结合位点区域结构变化对整体相互作用强弱的影响的实验数据，将有利于进一步分析非结合位点的重要性。

通过基于PDZ结构域的留一法验证了新模型的有效性，同时也指出新模型的性能与近邻PDZ结构域的序列相似度有很大的关联。这一个结果指导人们需要进一步地扩充PDZ结构域的序列空间，尽可能多地获取序列之间不是太相似的PDZ结构域的短肽相互作用数据。同时，为了使新模型能具有更有效地实现各个相近物种之间的PDZ结构域-短肽相互作用强弱的预测的性能，也需要获取更多的关于各个不同物种PDZ结构域的训练数据。

本章提出的新模型表明了整合负类相互作用数据能提高预测性能，未来一个有意义的工作是整合更多的来自不同数据源的相互作用数据，扩展新模型能实现既可整合负类样本点，也可有效利用正类样本点。例如，可以改进新模型，以使其可以整合来自噬菌体展示的正类相互作用对数据集，进而增强模型的普适性和提高模型的预测性能。

本章的结果表明基于“多结构域”的模型的预测性能优于基于“单结构域”的模型的预测性能，同时也表明“最近邻”的PDZ结构域的相关数据提供了主要的信息。未来一个需要进一步考察的有趣问题是，针对每一个测试PDZ结构域，如何选择有效的训练集进行构建预测模型。例如，根据PDZ结构域的短肽结合profile进行聚类，会发现已知实验数据的85个PDZ结构域可以形成若干个明显的类别。未来可以考虑根据不同的聚类类别构建预测模型，或者针对每一个测试PDZ结构域，可以根据PDZ结构域的序列相似度，选择相似度高于某设定阈值的结构域相关数据进行构建模型。设计不同的策略选取训练集并构建相应的预测模型，将是未来需要考虑的一个有趣的研究方向。

大部分的PDZ结构域通常与配体蛋白质的C端短肽序列相结合以行使其功能，但是近年来也有相关文献开始报道PDZ可以与配体蛋白质的其他内部模体、配体蛋白质中的PDZ结构域和Lipids相互作用[138]。与PDZ结构域相互结合的内部模体和C端短肽序列（模体）是否一致？在这两种不同的情形下，PDZ结构域序列的关键结合位点是否一致？PDZ结构域与Lipids结合的机理又是怎样？这一系列的问题将引导我们对PDZ结构域的相互作用特性进行更深入地研究。

第五章 结论与展望

论文的研究内容属于生物信息学、系统生物学的范畴，以生物学问题为驱动，以机器学习和运筹学理论为工具，以相互作用组学为研究对象，兼顾了蛋白质与核酸分子和蛋白质与蛋白质相互作用两大领域中的几个重要问题。DNA/RNA结合蛋白质是生物体内一类非常重要的蛋白质，本文的工作为生物学家研究这一类蛋白质提供了新的工具和新的视角。WW结构域和PDZ结构域是信号传导中两类非常重要的短肽识别结构域，与许多人类疾病有关。论文针对当前WW结构域和PDZ结构域-短肽相互作用研究面临的不同问题分别进行了系统的研究，加深了对其的认识，为生物学家开展进一步的同类研究工作提供了非常重要且丰富的资源。同时，本文在对生物学问题抽象成数学问题的过程中，给出了问题的新提法，并丰富了支持向量机的模型，反过来促进了支持向量机的发展和完善。本章对论文的主要工作做简要的总结，并就未来可能的工作做进一步的展望。

5.1 结论

论文从计算生物学的角度预测了与核酸大分子结合的功能蛋白质，并对蛋白质与蛋白质相互作用领域的结构域-短肽相互作用进行了深入的研究。具体地，主要有以下几个方面的成果。

（1）从蛋白质序列信息角度，应用描述蛋白质序列的新的特征编码方式——“三联体”编码方式对蛋白质进行了特征编码，并采用支持向量分类机模型，构建了DNA结合蛋白质分类机和RNA结合蛋白质分类机，实现对DNA和RNA结合蛋白质的预测。数值实验表明，对DNA结合蛋白质分类器而言，模型分类精度达到78.93%，比前人的工作有了较大的提高；而对RNA结合蛋白质也得到了与前人相当的分类性能。同时，还构建了区分DNA结合蛋白质与RNA结合蛋白质的DNA-RNA结合蛋白质分类机，交叉验证结果表明本文采用的特征能较好地区分DNA结合蛋白质和RNA结合蛋白质。本文采用“最大相关-最小冗余”特征选择方法选取了可以分别识别DNA结合蛋白质和RNA结合蛋白质的最为有效的部分“三联体”特征，并发现它们大部分都发生在DNA结合蛋白质或者RNA结合蛋白质的结合表面。这表明“三联体”特征有效地表征了DNA结合蛋白质或者RNA结合蛋白质在结合位点处的结合模式。此外，本文的研究结果还表明，基于Swiss-prot数据库的蛋白质全序列信息构建的分类器与基于PDB数据库中的蛋白质子序列或者蛋白质结构域序列信息构建的分类器之间不能有效地彼此进行预测，表明基于不同序列信息的分类器需要分别构建。

本文的工作是对前人工作的一个改进，通过分类其模型的构建和对“三联体”特征的分析，为科研人员进行DNA结合蛋白质和RNA结合蛋白质相关领域的研究提供了新的预测工具，并为人们认识DNA结合蛋白质和RNA结合蛋白质的结合特性提供了新的角度。

（2） 对人类蛋白质中WW结构域-短肽相互作用预测开展了深入的研究。本文采用“二值正交“编码方式分别对WW结构域和短肽序列进行特征编码，并应用两类支持向量分类机对WW结构域-短肽相互作用对是否发生相互作用进行了建模。本文的试验结果表明仅仅从WW结构域和短肽配体的序列信息预测它们之间的相互作用是行之有效的（针对每一个WW结构域的平均的

分类精度AUC值高达0.90）。本文尝试了多种不同的特征编码方式，发现仅利用氨基酸序列信息的分类模型达到了最佳的分类性能，这说明氨基酸序列信息本身富含了用于识别WW结构域-短肽相互作用的丰富信息。本文的研究还表明，整合不同的WW结构域的相互作用数据构建分类器比仅利用各自的相互作用数据或者最近邻的WW结构域的相互作用数据构建的分类器具有更佳的分类性能。此外，本文还通过采用支持向量顺序回归机模型，实现了对WW结构域-短肽相互作用强弱水平的预测，这给生物学家提供了更详尽的信息。

本文构建的预测模型是基于大规模实验数据训练得到的，率先实现了在整个人类蛋白质组层面上的针对WW结构域和其配体之间相互作用强弱水平的预测。本文的工作不仅给生物学家进行WW结构域-短肽相互作用研究提供了重要的工具和丰富的资源，也为科研人员进一步的研究提供了新的方向。

（3）针对PDZ结构域-短肽相互作用的定量化预测问题开展了研究工作。本文根据Stiffler和Chen等人实验测定的关于PDZ结构域-短肽相互作用的具有亲和力大小的“半量化”数据特性，设计了新的预测模型——半量化支持向量回归机，同时考虑了定量化的正类样本和定性化的负类样本，扩展了传统的仅能考虑定量化数据的支持向量回归模型。本文的结果表明了仅从PDZ结构域和短肽序列的氨基酸序列信息进行对它们之间的相互作用强弱的预测是可行的。同时也表明整合了负类样本信息的新模型能有效地提高传统回归模型的预测性能。本文的试验结果还表明了基于“多结构域”的模型的预测性能优于基于“单结构域”的模型的预测性能，同时也表明了“最近邻”的PDZ结构域的相互作用数据对预测模型提供了主要的信息。此外，本文的试验结果还表明了新模型能准确地预测对短肽序列进行单点变异之后的新短肽序列与PDZ结构域之间相互作用的亲和力的大小。本文提出的新模型不仅能实现对PDZ结构域-短肽序列之间是否发生相互作用的预测，也能对发生相互作用的PDZ结构域-短肽对预测其强弱（亲和力大小）。

本文的工作基于首次通过高通量实验获得的PDZ结构域-短肽相互作用亲和力的最新数据，构建了预测模型，实现了对PDZ结构域-短肽相互作用的定量化预测。本文的工作不仅给研究PDZ结构域–短肽之间的定量化相互作用提供了新的研究思路和新的预测工具，也为人们深刻认识PDZ结构域–短肽相互作用的内在机理提供了新的视角。同时还为人们从事其他结构域–短肽的定量化相互作用研究起到了非常有价值的借鉴作用。

5.2 展望

论文主要研究了蛋白质与核酸分子相互作用、蛋白质与蛋白质相互作用中的若干预测问题，取到了一定的结果，但是仍然有许多值得我们进一步思考的问题和改进的方面。具体的可以从以下几个方面进行考虑：

本文主要是从蛋白质序列的角度实现对蛋白质结构域-短肽相互作用的预测，未来可以考虑结合已知蛋白质的三维结构，以期可以构造出利于抓住结构域-短肽相互作用的结构特征，结合新的机器学习模型，提高对结构域-短肽相互作用的预测性能。此外，有相关文献表明结构域内部的loop区域对于结构域的识别特异性和结构域短肽相互作用的强弱均有显著影响[159, 160]，如何有效的编码结构域内部的loop区域将是未来工作的一个重要方向。另外，越来越多的文献表明蛋白质的disorder区在结构域-短肽相互作用中，尤其是在瞬时相互作用中起到非常重要的作用[161, 162]。

这些disorder区在生物功能上具有相当重要的作用，加入这部分信息也将有助于更好的构建预测结构域-短肽相互作用的模型。

目前蛋白质结构域-短肽相互作用预测越来越受到学者的关注，除了本文涉及到的PDZ和WW结构域之外，其中比较热点的还有SH2、SH3等结构域。开展针对这些结构域的结构域–短肽相互作用预测工作将是一件非常有意义而重要的课题。例如，SH2结构域主要识别磷酸化后的含有Y或者S、T位点的短肽序列，针对SH2结构域-短肽相互作用的计算模型已有发表[75, 163]，但是由于SH2结构域-短肽相互作用的预测较为困难，它们之间相互作用所受到的影响因素更多，如何有效地提高预测精度仍是一个重要的研究课题。

与蛋白质结构域-短肽相互作用密切相关的一类相互作用问题是药物-靶点相互作用。近年来，药物-靶点相互作用的研究已成为药物发现领域研究的重点。由于该领域关注的药物更多的是以化学结构形式出现，所以给研究者带来了新的挑战和机遇。日本学者Yamanishi 等人整合了化学分子结构数据、蛋白质序列数据和药理学数据，设计了不同的计算模型，实现了对药物-靶点相互作用的预测[164-166]。虽然，给定已知药物分子，预测靶点蛋白已经取得了一些成果，但是给定已知药理学上具有关键作用的蛋白质家族，实现对药物分子的识别仍然处于探索阶段[167]。加之药物化学分子的搜索空间巨大，从计算的角度预测最佳的药物-靶点相互作用对仍然任重而道远。

除了蛋白质与核酸分析相互作用、蛋白质与蛋白质相互作用外，近年来，也开展了其他分子之间的相互作用研究，如研究RNA与RNA之间的相互作用，包括微RNA（microRNA）与信使RNA（mRNA）之间的相互作用等。Ragan等人[168]提出了采用物理学中的热动力学模型预测RNA-RNA相互作用强弱，并对果蝇和人类中的实验数据进行了验证。但是利用机器学习模型来进行预测RNA-RNA相互作用并未见有报道，如何设计有效模型算法从序列的角度对RNA-RNA相互作用进行预测将是未来一个比较有趣的研究方向。

本文只是从相互作用组学的几个小方面开展了系统生物学的研究，相信随着人们对蛋白质功能及其与其他大分子之间的相互作用网络的深入研究，随着人们对系统生物学的深入研究，更多的生物现象和生物机理将会逐渐地被发现和确定，生物体内的奥妙也将会进一步得到揭示。

参考文献

[1] 孙啸,陆祖宏,谢建明, *生物信息学基础.* 清华大学出版社: 北京, 2005.

[2] http://en.wikipedia.org/wiki/amino_acid.

[3] Bartel D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell,* 2009, 136 (2), 215-233.

[4] http://en.wikipedia.org/wiki/Protein.

[5] http://en.wikipedia.org/wiki/Protein_domain.

[6] Finn R. D.,Tate J.,Mistry J.,Coggill P. C.,Sammut S. J.,Hotz H.-R.,Ceric G.,Forslund K.,Eddy S. R.,Sonnhammer E. L. L.,Bateman A. The Pfam protein families database. *Nucl. Acids Res.,* 2008, 36 (suppl_1), D281-288.

[7] Schultz J.,Milpetz F.,Bork P.,Ponting C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A,* 1998, 95 (11), 5857-5864.

[8] Marchler-Bauer A.,Lu S.,Anderson J. B.,Chitsaz F.,Derbyshire M. K.,DeWeese-Scott C.,Fong J. H.,Geer L. Y.,Geer R. C.,Gonzales N. R.,Gwadz M.,Hurwitz D. I.,Jackson J. D.,Ke Z.,Lanczycki C. J.,Lu F.,Marchler G. H.,Mullokandov M.,Omelchenko M. V.,Robertson C. L.,Song J. S.,Thanki N.,Yamashita R. A.,Zhang D.,Zhang N.,Zheng C.,Bryant S. H. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res,* 2011, 39 (Database issue), D225-229.

[9] Servant F.,Bru C.,Carrere S.,Courcelle E.,Gouzy J.,Peyruc D.,Kahn D. ProDom: automated clustering of homologous domains. *Brief Bioinform,* 2002, 3 (3), 246-251.

[10] Sanchez C.,Lachaize C.,Janody F.,Bellon B.,Roder L.,Euzenat J.,Rechenmann F.,Jacq B. Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Res,* 1999, 27 (1), 89-94.

[11] Kumar M. D.,Bava K. A.,Gromiha M. M.,Prabakaran P.,Kitajima K.,Uedaira H.,Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res,* 2006, 34 (Database issue), D204-206.

[12] Uetz P.,Giot L.,Cagney G.,Mansfield T. A.,Judson R. S.,Knight J. R.,Lockshon D.,Narayan V.,Srinivasan M.,Pochart P.,Qureshi-Emili A.,Li Y.,Godwin B.,Conover D.,Kalbfleisch T.,Vijayadamodar G.,Yang M. J.,Johnston M.,Fields S.,Rothberg J. M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature,* 2000, 403 (6770), 623-627.

[13] Ito T.,Chiba T.,Ozawa R.,Yoshida M.,Hattori M.,Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America,* 2001, 98 (8), 4569-4574.

[14] Rigaut G.,Shevchenko A.,Rutz B.,Wilm M.,Mann M.,Seraphin B. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology,* 1999, 17 (10), 1030-1032.

[15] Gavin A. C.,Bosche M.,Krause R.,Grandi P.,Marzioch M.,Bauer A.,Schultz J.,Rick J. M.,Michon A. M.,Cruciat C. M.,Remor M.,Hofert C.,Schelder M.,Brajenovic M.,Ruffner H.,Merino A.,Klein K.,Hudak M.,Dickson D.,Rudi T.,Gnau V.,Bauch A.,Bastuck S.,Huhse B.,Leutwein C.,Heurtier M. A.,Copley R. R.,Edelmann A.,Querfurth E.,Rybin V.,Drewes G.,Raida M.,Bouwmeester T.,Bork P.,Seraphin B.,Kuster B.,Neubauer G.,Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* 2002, 415 (6868), 141-147.

[16] Pawson T.,Nash P. Assembly of Cell Regulatory Systems Through Protein Interaction Domains. *Science,* 2003, 300 (5618), 445-452.

[17] Xenarios I.,Salwinski L.,Duan X. J.,Higney P.,Kim S. M.,Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res,* 2002, 30 (1), 303-305.

[18] Bader G. D.,Betel D.,Hogue C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res,* 2003, 31 (1), 248-250.

[19] Ceol A.,Chatr Aryamontri A.,Licata L.,Peluso D.,Briganti L.,Perfetto L.,Castagnoli L.,Cesareni G. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res,* 2010, 38 (Database issue), D532-539.

[20] Guldener U.,Munsterkotter M.,Oesterheld M.,Pagel P.,Ruepp A.,Mewes H. W.,Stumpflen V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res,* 2006, 34 (Database issue), D436-441.

[21] Pagel P.,Kovac S.,Oesterheld M.,Brauner B.,Dunger-Kaltenbach I.,Frishman G.,Montrone C.,Mark P.,Stumpflen V.,Mewes H. W.,Ruepp A.,Frishman D. The MIPS mammalian protein-protein interaction database. *Bioinformatics,* 2005, 21 (6), 832-834.

[22] Smialowski P.,Pagel P.,Wong P.,Brauner B.,Dunger I.,Fobo G.,Frishman G.,Montrone C.,Rattei T.,Frishman D.,Ruepp A. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res,* 2010, 38 (Database issue), D540-544.

[23] Stark C.,Breitkreutz B. J.,Reguly T.,Boucher L.,Breitkreutz A.,Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res,* 2006, 34 (Database issue), D535-539.

[24] Brown K. R.,Jurisica I. Online predicted human interaction database. *Bioinformatics,* 2005, 21 (9), 2076-2082.

[25] Keskin O.,Nussinov R.,Gursoy A. PRISM: protein-protein interaction prediction by structural matching. *Methods Mol Biol,* 2008, 484, 505-521.

[26] Chen Y. C.,Lo Y. S.,Hsu W. C.,Yang J. M. 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res,* 2007, 35 (Web Server issue), W561-567.

[27] Finn R. D.,Marshall M.,Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics,* 2005, 21 (3), 410-412.

[28] Stein A.,Russell R. B.,Aloy P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res,* 2005, 33 (Database issue), D413-417.

[29] Raghavachari B.,Tasneem A.,Przytycka T. M.,Jothi R. DOMINE: a database of protein domain interactions. *Nucleic Acids Res,* 2008, 36 (Database issue), D656-661.

[30] Yellaboina S.,Tasneem A.,Zaykin D. V.,Raghavachari B.,Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res,* 2011, 39 (Database issue), D730-735.

[31] Luo Q.,Pagel P.,Vilne B.,Frishman D. DIMA 3.0: Domain Interaction Map. *Nucleic Acids Res,* 2011, 39 (Database issue), D724-729.

[32] Ng S. K.,Zhang Z.,Tan S. H.,Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res,* 2003, 31 (1), 251-254.

[33] Gould C. M.,Diella F.,Via A.,Puntervoll P.,Gemund C.,Chabanis-Davidson S.,Michael S.,Sayadi A.,Bryne J. C.,Chica C.,Seiler M.,Davey N. E.,Haslam N.,Weatheritt R. J.,Budd A.,Hughes T.,Pas

J.,Rychlewski L.,Trave G.,Aasland R.,Helmer-Citterich M.,Linding R.,Gibson T. J. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res,* 2010, 38 (Database issue), D167-180.

[34] Ceol A.,Chatr-Aryamontri A.,Santonico E.,Sacco R.,Castagnoli L.,Cesareni G. DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res,* 2007, 35, D557-D560.

[35] Encinar J. A.,Fernandez-Ballester G.,Sanchez I. E.,Hurtado-Gomez E.,Stricher F.,Beltrao P.,Serrano L. ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics,* 2009, 25 (18), 2418-2424.

[36] Vanhee P.,Reumers J.,Stricher F.,Baeten L.,Serrano L.,Schymkowitz J.,Rousseau F. PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res,* 2010, 38 (Database issue), D545-551.

[37] Dinkel H.,Chica C.,Via A.,Gould C. M.,Jensen L. J.,Gibson T. J.,Diella F. Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res,* 2011, 39 (Database issue), D261-267.

[38] Beuming T.,Skrabanek L.,Niv M. Y.,Mukherjee P.,Weinstein H. PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics,* 2005, 21 (6), 827-828.

[39] Deng N.,Tian Y.,Zhang C., *Support Vector Machines: Theory, Algorithms, and Extensions.* CRC Press: 2012. (In Press).

[40] 邓乃扬,田英杰, *支持向量机：理论、算法与拓展.* 科学出版社: 北京, 2009.

[41] Vapnik V. N., *The nature of statistical learning theory*. Springer: New York, 1995.

[42] 邓乃扬,田英杰, *数据挖掘中的新方法--支持向量机.* 科学出版社: 北京, 2004.

[43] Smola A. J.,Scholkopf B. A tutorial on support vector regression. *Stat Comput,* 2004, 14 (3), 199-222.

[44] Hsu C. W.,Lin C. J. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks,* 2002, 13 (2), 415-425.

[45] Bennett K. P.,Demiriz A., Semi-Supervised Support Vector Machines. In *Advances in Proceedings of Neural Information Processing Systems*, Kearns, M. S.; Solla, S. A.; Cohn, D. A., Eds. MIT Press: Cambridge, MA, 1998; Vol. 12, pp 368-374.

[46] Shashua A.,Levin A., Ranking with Large Margin Principle: Two Approaches. In *Advances in Neural Information Processing Systems*, 2002; Vol. 15, pp 937–944.

[47] Chu W.,Keerthi S. S., New Approaches to Support Vector Ordinal Regression. In *Proceedings of the 22 nd International Conference on Machine Learning*, Bonn, Germany, 2005.

[48] Chu W.,Keerthi S. S. Support vector ordinal regression. *Neural Comput,* 2007, 19 (3), 792-815.

[49] 杨志霞. 支持向量顺序回归机和多分类问题的研究，[博士学位论文]. 中国农业大学, 北京, 2007.

[50] Eisen J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res,* 1998, 8 (3), 163-167.

[51] Teichmann S. A.,Murzin A. G.,Chothia C. Determination of protein function, evolution and interactions by structural genomics. *Current Opinion in Structural Biology,* 2001, 11 (3), 354-363.

[52] Cai Y. D.,Lin S. L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta,* 2003, 1648 (1-2), 127-133.

[53] Yu X.,Cao J.,Cai Y.,Shi T.,Li Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J Theor Biol,* 2006, 240 (2), 175-184.

[54] Han L. Y.,Cai C. Z.,Lo S. L.,Chung M. C.,Chen Y. Z. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA,* 2004, 10 (3), 355-368.

[55] Ahmad S.,Sarai A. Moment-based Prediction of DNA-binding Proteins. *J Mol Biol,* 2004, 341 (1), 65-71.

[56] Bhardwaj N.,Langlois R. E.,Zhao G.,Lu H. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res,* 2005, 33 (20), 6486-6493.

[57] Pellegrini M.,Marcotte E. M.,Thompson M. J.,Eisenberg D.,Yeates T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America,* 1999, 96 (8), 4285-4288.

[58] Marcotte E. M.,Pellegrini M.,Ng H. L.,Rice D. W.,Yeates T. O.,Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science,* 1999, 285 (5428), 751-753.

[59] Dandekar T.,Snel B.,Huynen M.,Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences,* 1998, 23 (9), 324-328.

[60] Sprinzak E.,Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol,* 2001, 311 (4), 681-692.

[61] Deng M. H.,Mehta S.,Sun F. Z.,Chen T. Inferring domain-domain interactions from protein-protein interactions. *Genome Research,* 2002, 12 (10), 1540-1548.

[62] Riley R.,Lee C.,Sabatti C.,Eisenberg D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biology,* 2005, 6 (10), -.

[63] Nye T. M. W.,Berzuini C.,Gilks W. R.,Babu M. M.,Teichmann S. A. Statistical analysis of domains in interacting protein pairs. *Bioinformatics,* 2005, 21 (7), 993-1001.

[64] Guimaraes K. S.,Jothi R.,Zotenko E.,Przytycka T. M. Predicting domain-domain interactions using a parsimony approach. *Genome Biology,* 2006, 7 (11), -.

[65] Bock J. R.,Gough D. A. Predicting protein-protein interactions from primary structure. *Bioinformatics,* 2001, 17 (5), 455-460.

[66] Chen X. W.,Liu M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics,* 2005, 21 (24), 4394-4400.

[67] Shen J.,Zhang J.,Luo X.,Zhu W.,Yu K.,Chen K.,Li Y.,Jiang H. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A,* 2007, 104 (11), 4337-4341.

[68] Hubbard S. R.,Bergamin E.,Wu J. H. Structural basis for phosphotyrosine recognition by suppressor of cytokine signaling-3. *Structure,* 2006, 14 (8), 1285-1292.

[69] Brannetti B.,Via A.,Cestra G.,Cesareni G.,Citterich M. H. SH3-SPOT: An algorithm to predict preferred ligands to different members of the SH3 gene family. *J Mol Biol,* 2000, 298 (2), 313-328.

[70] Wollacott A. M.,Desjarlais J. R. Virtual interaction profiles of proteins. *J Mol Biol,* 2001, 313 (2), 317-342.

[71] Fernandez-Ballester G.,Beltrao P.,Gonzalez J. M.,Song Y. H.,Wilmanns M.,Valencia A.,Serrano L. Structure-based prediction of the Saccharomyces cerevisiae SH3-ligand interactions. *J Mol Biol,* 2009, 388 (4), 902-916.

[72] Hou T.,Xu Z.,Zhang W.,McLaughlin W. A.,Case D. A.,Xu Y.,Wang W. Characterization of Domain-Peptide Interaction Interface: A Generic Structure-based Model to Decipher the Binding Specificity of SH3 Domains. *Mol Cell Proteomics,* 2009, 8 (4), 639-649.

[73] Ferraro E.,Via A.,Ausiello G.,Helmer-Citterich M. A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics,* 2006, 22 (19), 2333-2339.

[74] Zhang L.,Shao C.,Zheng D.,Gao Y. An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. *Mol Cell Proteomics,* 2006, 5 (7), 1224-1232.

[75] Wunderlich Z.,Mirny L. A. Using genome-wide measurements for computational prediction of SH2-peptide interactions. *Nucleic Acids Res,* 2009, 37 (14), 4629-4641.

[76] Lehrach W. P.,Husmeier D.,Williams C. K. A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics,* 2006, 22 (5), 532-540.

[77] Obenauer J. C.,Cantley L. C.,Yaffe M. B. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research,* 2003, 31 (13), 3635-3641.

[78] Reiss D. J.,Schwikowski B. Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics,* 2004, 20(Suppl 1), i274–i282.

[79] Wiedemann U.,Boisguerin P.,Leben R.,Leitner D.,Krause G.,Moelling K.,Volkmer-Engert R.,Oschkinat H. Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol,* 2004, 343 (3), 703-718.

[80] McLaughlin W. A.,Hou T. J.,Wang W. Prediction of binding sites of peptide recognition domains: An application on Grb2 and SAP SH2 domains. *J Mol Biol,* 2006, 357 (4), 1322-1334.

[81] Reche P. A.,Glutting J. P.,Reinherz E. L. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol,* 2002, 63 (9), 701-709.

[82] Zhang H.,Lundegaard C.,Nielsen M. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics,* 2009, 25 (1), 83-89.

[83] Donnes P.,Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics,* 2002, 3.

[84] Jacob L.,Vert J.-P. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics,* 2008, 24 (3), 358-366.

[85] Liu W.,Meng X.,Xu Q.,Flower D.,Li T. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics,* 2006, 7 (1), 182.

[86] Nielsen M.,Lundegaard C.,Blicher T.,Lamberth K.,Harndahl M.,Justesen S.,Roder G.,Peters B.,Sette A.,Lund O.,Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One,* 2007, 2 (8), e796.

[87] Otte L.,Wiedemann U.,Schlegel B.,Pires J. R.,Beyermann M.,Schmieder P.,Krause G.,Volkmer-Engert R.,Schneider-Mergener J.,Oschkinat H. WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. *Protein Sci,* 2003, 12 (3), 491-500.

[88] Hu H.,Columbus J.,Zhang Y.,Wu D.,Lian L.,Yang S.,Goodwin J.,Luczak C.,Carter M.,Chen L.,James M.,Davis R.,Sudol M.,Rodwell J.,Herrero J. J. A map of WW domain family interactions. *Proteomics,* 2004, 4 (3), 643-655.

[89] Ingham R. J.,Colwill K.,Howard C.,Dettwiler S.,Lim C. S. H.,Yu J.,Hersi K.,Raaijmakers J.,Gish G.,Mbamalu G.,Taylor L.,Yeung B.,Vassilovski G.,Amin M.,Chen F.,Matskova L.,Winberg G.,Ernberg I.,Linding R.,O'donnell P.,Starostine A.,Keller W.,Metalnikov P.,Stark C.,Pawson T.

WW domains provide a platform for the assembly of multiprotein networks. *Mol Cell Biol,* 2005, 25 (16), 7092-7106.

[90] Hesselberth J.,Miller J.,Golob A.,Stajich… J. Comparative analysis of Saccharomyces cerevisiae WW domains and their interacting proteins. *Genome Biology,* 2006.

[91] Schleinkofer K.,Wiedemann U.,Otte L.,Wang T.,Krause G.,Oschkinat H.,Wade R. C. Comparative structural and energetic analysis of WW domain-peptide interactions. *J Mol Biol,* 2004, 344 (3), 865-881.

[92] Tonikian R.,Zhang Y.,Sazinsky S. L.,Currell B.,Yeh J.-H.,Reva B.,Held H. A.,Appleton B. A.,Evangelista M.,Wu Y.,Xin X.,Chan A. C.,Seshagiri S.,Lasky L. A.,Sander C.,Boone C.,Bader G. D.,Sidhu S. S. A Specificity Map for the PDZ Domain Family. *PLoS Biol,* 2008, 6 (9), e239.

[93] Hui S.,Bader G. D. Proteome scanning to predict PDZ domain interactions using support vector machines. *BMC Bioinformatics,* 2010, 11, 507.

[94] Stiffler M. A.,Chen J. R.,Grantcharova V. P.,Lei Y.,Fuchs D.,Allen J. E.,Zaslavskaia L. A.,MacBeath G. PDZ domain binding selectivity is optimized across the mouse proteome. *Science,* 2007, 317 (5836), 364-369.

[95] Chen J. R.,Chang B. H.,Allen J. E.,Stiffler M. A.,MacBeath G. Predicting PDZ domain-peptide interactions from primary sequences. *Nat Biotech,* 2008, 26 (9), 1041-1045.

[96] Dervan P. Design of sequence-specific DNA-binding molecules. *Science,* 1986, 232 (4749), 464-471.

[97] Travers A. A.,Buckle M., *DNA-protein interactions: A practical approach*. Oxford University Press: Usa, 2000.

[98] Lunde B. M.,Moore C.,Varani G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol,* 2007, 8 (6), 479-490.

[99] 李英贤,贺福初. DNA与蛋白质相互作用研究方法. *生命的化学,* 2003, 23 (4), 306-308.

[100] Bork P.,Koonin E. V. Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet,* 1998, 18 (4), 313-318.

[101] Luscombe N. M.,Austin S. E.,Berman H. M.,Thornton J. M., An overview of the structures of protein-DNA complexes. In *Genome Biol*, 2000; Vol. 1, p REVIEWS001.

[102] Langlois R. E.,Carson M. B.,Bhardwaj N.,Lu H., Learning to translate sequence and structure to function: identifying DNA binding and membrane binding proteins. In *Ann Biomed Eng*, 2007; Vol. 35, pp 1043-1052.

[103] Ahmad S.,Gromiha M. M.,Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics,* 2004, 20 (4), 477-486.

[104] Stawiski E. W.,Gregoret L. M.,Mandel-Gutfreund Y. Annotating Nucleic Acid-Binding Function Based on Protein Structure. *J Mol Biol,* 2003, 326 (4), 1065-1079.

[105] Shanahan H. P. Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res,* 2004, 32 (16), 4732-4741.

[106] Shazman S.,Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput Biol,* 2008, 4 (8), e1000146.

[107] Boeckmann B.,Bairoch A.,Apweiler R.,Blatter M. C.,Estreicher A.,Gasteiger E.,Martin M. J.,Michoud K.,O'Donovan C.,Phan I.,Pilbout S.,Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res,* 2003, 31 (1), 365-370.

[108] Shao X.,Tian Y.,Wu L.,Wang Y.,Jing L.,Deng N. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *J Theor Biol,* 2009, 258 (2), 289-293.

[109] Li W.,Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 2006, 22 (13), 1658-1659.

[110] Berman H. M.,Westbrook J.,Feng Z.,Gilliland G.,Bhat T. N.,Weissig H.,Shindyalov I. N.,Bourne P. E. The Protein Data Bank. *Nucleic Acids Res,* 2000, 28 (1), 235-242.

[111] Kumar M.,Gromiha M. M.,Raghava G. P. S. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics,* 2007, 8, 463.

[112] Li Z. R.,Lin H. H.,Han L. Y.,Jiang L.,Chen X.,Chen Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucl. Acids Res.,* 2006, 34 (suppl_2), W32-37.

[113] Chang C.-C.,Lin C.-J. LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm 2001.

[114] Ding C.,Peng H. Minimum redundancy feature selection from microarray gene expression data. *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE,* 2003, 523 - 528.

[115] Dong D.,Shao X.,Deng N.,Zhang Z. Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res,* 2011, 39 (2), 403-413.

[116] Baldi P.,Brunak S.,Chauvin Y.,Andersen C. A.,Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 2000, 16 (5), 412-424.

[117] Staker B. L.,Korber P.,Bardwell J. C.,Saper M. A. Structure of Hsp15 reveals a novel RNA-binding motif. *EMBO J,* 2000, 19 (4), 749-757.

[118] Lewis B. A.,Walia R. R.,Terribilini M.,Ferguson J.,Zheng C.,Honavar V.,Dobbs D. PRIDB: a Protein-RNA Interface Database. *Nucleic Acids Res,* 2011, 39 (Database issue), D277-282.

[119] Rokach L. Ensemble-based classifiers. *Artificial Intelligence Review,* 2010, 33 (1), 1-39.

[120] Wu J.,Liu H.,Duan X.,Ding Y.,Wu H.,Bai Y.,Sun X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics,* 2009, 25 (1), 30-35.

[121] Wang L. J.,Brown S. J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res,* 2006, 34, W243-W248.

[122] Terribilini M.,Lee J. H.,Yan C. H.,Jernigan R. L.,Honavar V.,Dobbs D. Prediction of RNA binding sites in proteins from amino acid sequence. *Rna-a Publication of the Rna Society,* 2006, 12 (8), 1450-1462.

[123] Wang L. J.,Yang M. Q.,Yang J. Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics,* 2009, 10.

[124] Ofran Y.,Mysore V.,Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics,* 2007, 23 (13), I347-I353.

[125] Liu Z. P.,Wu L. Y.,Wang Y.,Zhang X. S.,Chen L. N. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics,* 2010, 26 (13), 1616-1622.

[126] Bullock A. N.,Fersht A. Rescuing the function of mutant p53. *Nat Rev Cancer,* 2001, 1 (1), 68-76.

[127] Ho S. Y.,Yu F. C.,Chang C. Y.,Huang H. L. Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. *Biosystems,* 2007, 90 (1), 234-241.

[128] Bulyk M. L.,Berger M. F.,Badis G.,Gehrke A. R.,Talukder S.,Philippakis A. A.,Pena-Castillo L.,Alleyne T. M.,Mnaimneh S.,Botvinnik O. B.,Chan E. T.,Khalid F.,Zhang W.,Newburger

D.,Jaeger S. A.,Morris Q. D.,Hughes T. R. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell,* 2008, 133 (7), 1266-1276.

[129] Alleyne T. M.,Pena-Castillo L.,Badis G.,Talukder S.,Berger M. F.,Gehrke A. R.,Philippakis A. A.,Bulyk M. L.,Morris Q. D.,Hughes T. R. Predicting the binding preference of transcription factors to individual DNA k-mers. *Bioinformatics,* 2009, 25 (8), 1012-1018.

[130] Cook K. B.,Kazan H.,Zuberi K.,Morris Q.,Hughes T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res,* 2011, 39 (Database issue), D301-308.

[131] Bork P.,Sudol M. The WW domain: a signalling site in dystrophin? *Trends Biochem Sci,* 1994, 19 (12), 531-533.

[132] Letunic I.,Doerks T.,Bork P. SMART 6: recent updates and new developments. *Nucleic Acids Res,* 2009, 37 (Database issue), D229-232.

[133] Macias M. J.,Gervais V.,Civera C.,Oschkinat H. Structural analysis of WW domains and design of a WW prototype. *Nat Struct Biol,* 2000, 7 (5), 375-379.

[134] Sudol M.,Hunter T. NeW wrinkles for an old domain. *Cell,* 2000, 103 (7), 1001-1004.

[135] Schleinkofer K.,Wiedemann U.,Otte L.,Wang T.,Krause G.,Oschkinat H.,Wade R. C., Comparative structural and energetic analysis of WW domain-peptide interactions. In *J Mol Biol*, 2004; Vol. 344, pp 865-881.

[136] Katoh K.,Kuma K.,Toh H.,Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res,* 2005, 33 (2), 511-518.

[137] Chen H. I.,Einbond A.,Kwak S. J.,Linn H.,Koepf E.,Peterson S.,Kelly J. W.,Sudol M. Characterization of the WW domain of human yes-associated protein and its polyproline-containing ligands. *J Biol Chem,* 1997, 272 (27), 17070-17077.

[138] Nourry C.,Grant S. G.,Borg J. P. PDZ domain proteins: plug and play! *Sci STKE,* 2003, 2003 (179), RE7.

[139] Dev K. K. Making protein interactions druggable: targeting PDZ domains. *Nature Reviews Drug Discovery,* 2004, 3 (12), 1047.

[140] Kim E.,Sheng M. PDZ domain proteins of synapses. *Nat Rev Neurosci,* 2004, 5 (10), 771-781.

[141] Stiffler M. A.,Grantcharova V. P.,Sevecka M.,MacBeath G. Uncovering Quantitative Protein Interaction Networks for Mouse PDZ Domains Using Protein Microarrays. *Journal of the American Chemical Society,* 2006, 128 (17), 5913-5922.

[142] Eo H. S.,Kim S.,Koo H.,Kim W. A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins. *Mol Cells,* 2009, 27 (6), 629-634.

[143] Chechik G.,Oh E.,Rando O.,Weissman J.,Regev A.,Koller D. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat Biotech,* 2008, 26 (11), 1251-1259.

[144] Lew E. D.,Furdui C. M.,Anderson K. S.,Schlessinger J. The Precise Sequence of FGF Receptor Autophosphorylation Is Kinetically Driven and Is Disrupted by Oncogenic Mutations. *Sci. Signal.,* 2009, 2 (58), ra6-.

[145] Zarrinpar A.,Park S.-H.,Lim W. A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature,* 2003, 426 (6967), 676-680.

[146] Marles J. A.,Dahesh S.,Haynes J.,Andrews B. J.,Davidson A. R. Protein-Protein Interaction Affinity Plays a Crucial Role in Controlling the Sho1p-Mediated Signal Transduction Pathway in Yeast. *Molecular Cell,* 2004, 14 (6), 813-823.

[147] Sette A.,Buus S.,Appella E.,Smith J. A.,Chesnut R.,Miles C.,Colon S. M.,Grey H. M. Prediction of Major Histocompatibility Complex Binding Regions of Protein Antigens by Sequence Pattern-Analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 1989, 86 (9), 3296-3300.

[148] Rammensee H. G.,Bachmann J.,Emmerich N. P. N.,Bachor O. A.,Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics,* 1999, 50 (3-4), 213-219.

[149] Sette A.,Fleri W.,Peters B.,Sathiamurthy M.,Bui H.-H.,Wilson S. A Roadmap for the Immunomics of Category A-C Pathogens. *Immunity,* 2005, 22 (2), 155-161.

[150] Bhasin M.,Singh H.,Raghava G. P. S. MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics,* 2003, 19 (5), 665-666.

[151] Nielsen M.,Lundegaard C.,Blicher T.,Peters B.,Sette A.,Justesen S.,Buus S.,Lund O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput Biol,* 2008, 4 (7), e1000107.

[152] Mangasarian O. L.,Shavlik J. W.,Wild E. W. Knowledge-Based Kernel Approximation. *J. Mach. Learn. Res.,* 2004, 5, 1127-1141.

[153] Mangasarian O. L.,Wild E. W. Nonlinear knowledge in kernel approximation. *IEEE Transactions on Neural Networks,* 2007, 18, 300-306.

[154] Mangasarian O. L.,David R. M. Large Scale Kernel Regression via Linear Programming. *Mach. Learn.,* 2002, 46 (1-3), 255-269.

[155] Sandberg M.,Eriksson L.,Jonsson J.,Sjostrom M.,Wold S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *Journal of Medicinal Chemistry,* 1998, 41 (14), 2481-2491.

[156] William R. Atchley J. Z., Andrew D. Fernandes, and Tanja Drüke Solving the protein sequence metric problem. *PNAS,* 2005, 102, 6395-6400.

[157] Bloom S. A. Similarity Indices in Community Studies: Potential Pitfalls. *Marine Ecology Progress Series,* 1981, 5, 125–128.

[158] Lockless S. W.,Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* 1999, 286 (5438), 295-299.

[159] Kaneko T.,Huang H.,Zhao B.,Li L.,Liu H.,Voss C. K.,Wu C.,Schiller M. R.,Li S. S.-C. Loops Govern SH2 Domain Specificity by Controlling Access to Binding Pockets. *Sci Signal,* 2010, 3 (120), ra34.

[160] Peng T.,Zintsmaster J. S.,Namanja A. T.,Peng J. W. Sequence-specific dynamics modulate recognition specificity in WW domains. *Nat Struct Mol Biol,* 2007, 14 (4), 325-331.

[161] Stein A.,Pache R. A.,Bernado P.,Pons M.,Aloy P. Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J,* 2009, 276 (19), 5390-5405.

[162] Perkins J. R.,Diboun I.,Dessailly B. H.,Lees J. G.,Orengo C. Transient protein-protein interactions: structural, functional, and network properties. *Structure,* 2010, 18 (10), 1233-1243.

[163] Sanchez I. E.,Beltrao P.,Stricher F.,Schymkowitz J.,Ferkinghoff-Borg J.,Rousseau F.,Serrano L. Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS Comput Biol,* 2008, 4 (4), e1000052.

[164] Yamanishi Y.,Araki M.,Gutteridge A.,Honda W.,Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics,* 2008, 24 (13), i232-240.

[165] Yamanishi Y.,Kotera M.,Kanehisa M.,Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics,* 2010, 26 (12), i246-254.

[166] Kevin B.,Yoshihiro Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics,* 2009, 25 (18), 2397-2403.

[167] Yabuuchi H.,Niijima S.,Takematsu H.,Ida T.,Hirokawa T.,Hara T.,Ogawa T.,Minowa Y.,Tsujimoto G.,Okuno Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol Syst Biol,* 2011, 7, 472.

[168] Ragan C.,Zuker M.,Ragan M. A. Quantitative Prediction of miRNA-mRNA Interaction Based on Equilibrium Concentrations. *PLoS Comput Biol,* 2011, 7 (2).

附录

A. 基于不同特征编码的半量化支持向量回归机模型之性能比较（Pearson相关系数）

Pearson	38 pairs	BSs-16AA (pairwise)		WholePDZ-118AA (pairwise)		Profeat+11f	Profeat+5f	Profeat+zscale	BS16AA +10AA	BS16AA + 5AA
PDZ domain name	Linear	Linear	Polynomial	Linear	**Polynomial**	RBF	RBF	RBF	Sparse20 RBF	Sparse20 RBF
CHAPSYN-110_2/3	0.927	0.955	0.942	0.933	0.936	0.688	0.752	0.754	0.942	0.934
CHAPSYN-110_3/3	0.565	0.674	0.933	0.705	0.874	0.802	0.691	0.710	0.933	0.895
GM1582_2/3	0.354	0.563	0.479	0.609	0.580	0.188	0.372	0.181	0.479	0.524
HTRA3_1/1	0.356	0.472	0.584	0.486	0.646	0.638	0.718	0.434	0.584	0.051
LIN7C_1/1	0.558	0.512	0.574	0.581	0.682	0.780	0.459	0.391	0.574	0.545
MAGI-2_2/6	0.779	0.741	0.761	0.836	0.769	0.423	0.597	0.583	0.761	0.862
MAGI-2_6/6	0.523	0.652	0.687	0.411	0.689	0.772	0.858	0.729	0.687	-0.466
MAGI-3_1/5	0.679	0.647	0.828	0.663	0.876	0.631	0.481	0.447	0.828	0.785
MALS2_1/1	0.374	0.772	0.722	0.631	0.612	0.341	0.555	0.400	0.722	0.677
OMP25_1/1	0.509	0.227	0.550	0.293	0.504	0.609	0.421	0.443	0.550	0.577
PDZK3_1/1	0.021	0.072	-0.074	-0.007	0.039	-0.427	-0.266	-0.489	-0.074	-0.426
PDZ-RGS3_1/1	-0.050	0.338	-0.065	0.143	0.027	0.076	0.563	0.578	-0.065	-0.112
PSD95_2/3	0.869	0.916	0.926	0.903	0.917	0.867	0.848	0.806	0.926	0.895
PSD95_3/3	0.676	0.756	0.781	0.856	0.880	0.669	0.730	0.834	0.781	0.890
PTP-BL_2/5	0.529	0.557	0.640	0.548	0.401	0.114	0.522	0.260	0.640	0.694
SAP102_2/3	0.924	0.800	0.958	0.938	0.938	0.660	0.688	0.716	0.958	0.954
SAP97_1/3	0.625	0.605	0.682	0.713	0.755	0.275	0.631	0.303	0.682	0.529
SAP97_2/3	0.915	0.951	0.913	0.935	0.949	0.872	0.879	0.885	0.913	0.937
SCRB1_3/4	0.474	0.598	0.712	0.677	0.694	0.548	0.404	0.479	0.712	0.620
SHANK1_1/1	0.442	0.814	0.969	0.981	0.980	0.832	0.615	0.504	0.969	0.969
SHANK3_1/1	0.909	0.733	0.310	0.577	0.509	0.217	0.623	0.645	0.310	0.590
G1-SYNTROPHIN_1/1	0.160	0.286	0.025	0.240	0.129	0.302	0.264	0.471	0.025	0.252
ZO-1_1/3	0.642	0.153	0.793	0.286	0.646	0.562	-0.169	0.550	0.793	0.224
Average Performance	0.555	0.600	0.636	0.606	**0.653**	0.497	0.532	0.505	0.636	0.539

致谢

值此论文完成之际，我首先感谢我的导师邓乃扬教授！

本论文从选题到完稿都是在导师邓乃扬教授的悉心指导下进行的。时光飞逝，五年的博士生涯即将结束，回想过往，邓老师不管是在学业、科研还是在生活等方面都给予了我许多关怀和帮助。邓老师渊博的知识体系，严谨的治学态度、缜密的思维，敏锐的洞察力，和忘我的敬业精神，使我耳濡目染，受益匪浅。邓老师不仅教给我如何做学问，更重要的是教导我如何做人。非常感谢邓老师指引我进入生物信息学领域，才有了我现在的博士论文课题。感谢邓老师支持我申请国家留学基金委的公派留学项目。自己在学习和科研上取得的每一点进步和成绩，都凝聚了邓老师您的教诲、付出和心血。在论文完成之际，对邓老师再次表示感谢。

我还要衷心地感谢多伦多大学的联合培养导师Gary Bader教授！在2008年9月至2010年9月的两年联合培养学习期间，您给了我非常大的指导和帮助，是生活和学习上的良师益友。您开阔的视野和对科学研究一丝不苟的态度深深感染着我。特别感谢您一直以来对我的包容和鼓励，这是我继续前行的勇气和动力。感谢您对我论文的选题和写作过程中给予的细心指导和严格要求。在此，谨向您表示最高的谢意。

衷心感谢北京理工大学的刘宝光教授、中国科学院的田英杰副研究员和新疆大学杨志霞副教授，他们对我的论文提出了宝贵的意见和建议。感谢中国农业大学王来生教授和经玲教授，他们在我读博期间给了我谆谆教诲和热心的帮助，给我的论文提出了宝贵的修改意见。感谢中国科学院的赖炎连研究员和我的硕士生导师贺国平教授，他们一直关心我的博士学业。

感谢在校和毕业的同门师兄弟、师姐妹们：赵琨、秦如新、王永翠、邵元海、徐岩、高婷婷、赵艳梅、王晓波、马新港、李玉欣等，是你们陪伴我度过了博士生涯，在生活上和学习上都给予了我非常大的帮助。特别感谢中国农业大学谭俊艳博士，对论文的修改提出了许多非常具体而宝贵的意见。

还要特别感谢中国科学院数学与系统科学院的Zhangroup讨论班。在讨论班上学习到的知识对我的博士论文研究非常有帮助。尤其要感谢吴凌云副研究员和王勇副研究员，他们在我进入生物信息学领域研究的过程中给予了很大的帮助。特别感谢张世华助理研究员对论文的写作提出了宝贵的修改意见。

诚挚地感谢多伦多大学Bader实验室的Chris Tan、David Gfeller、Shirley Hui及其他成员。感谢多伦多大学Zhang实验室的董东博士。与你们的讨论让我深受启发。感谢多伦多大学其他帮助过我的同学和朋友。

感谢我的女朋友刘伟伟，谢谢你一直以来对我的理解和陪伴。

最后，要特别感谢我的父亲和母亲！正是您们的不辞劳苦、辛勤付出、对我的理解和包容，才使我有勇气和信心完成博士学业。在未来的工作上，我会加倍努力，以报答您们的养育之恩！

谨以此文献给我最深爱的父亲和母亲。

个人简历

邵小健，男，浙江兰溪市人，1980年11月出生。1998年考入山东科技大学信息学院（获理学学士学位）。2002年考入山东科技大学信息学院（获理学硕士学位）。2006年考入中国农业大学理学院攻读管理学博士学位。2008年9月至2010年9月在加拿大多伦多大学进行博士联合培养学习。

在校期间发表论文

1. **Xiaojian Shao***, Chris Tan*, Courtney Voss, Shawn S. C. Li, Naiyang Deng, Gary D. Bader. A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. Bioinformatics 2010, 27(3): 383-390. doi:10.1093/bioinformatics/btq657. (SCI: IF = 4.92. ***-co-first author**)
2. Dong Dong*, **Xiaojian Shao***, Naiyang Deng and Zhaolei Zhang. Gene expression variations are proxies for stochastic noise. Nuclear Acid Research 2010, 39(2): 403-413. doi:10.1093/nar/gkq844. (SCI: IF = 7.479. ***-co-first author**)
3. Dong Dong, **Xiaojian Shao** and Zhaolei Zhang. Different effects of chromatin regulators and transcription factors on gene regulation: A nucleosomal perspective. Bioinformatics 2010, 27(2): 147-152. doi:10.1093/bioinformatics/btq637. (SCI: IF = 4.92)
4. Xiaobo Wang, Yongcui Wang, Yingjie Tian, **Xiaojian Shao**, Ling-Yun Wu, Naiyang Deng. Prediction of posttranslational modification sites from sequences with kernel methods. Journal of Computational Chemistry 2010, doi: 10.1002/jcc.21526. (SCI: IF = 3.769)
5. **Xiaojian Shao**, Yingjie Tian, Lingyun Wu, Yong Wang, Ling Jing and Naiyang Deng. Predicting DNA- and RNA-binding Proteins from Sequences with Kernel Methods, Journal of Theoretical Biology 2009, Vol. 258, NO. 2, pp. 289-293. (SCI: IF = 2.574)
6. Hua Duan, **Xiaojian Shao**, Weizhen Hou, Guoping He, Qingtian Zeng. An Incremental Learning Algorithm for Lagrangian Support Vector Machines. Pattern Recognition Letters 2009, Vol 30, Issue 15: 1384-1391.(SCI: IF = 1.59)
7. Yanzhong Liu, **Xiaojian Shao**, Xuhong Li, Xuehui Gao. The Study on a Real-time Forecasting Model for Short-term Traffic Flow Based on Online Incremental LSVR. The 6th International Conference Traffic & Transportation Studies 2008, pp. 852-861. (EI)
8. Tingting Gao, Yingjie Tian, **Xiaojian Shao**, Naiyang Deng. Accurate Prediction of Translation Initiation Sites by Universum SVM. In Proceedings of 2st International Symposium on Optimization and Systems Biology, Lecture Notes in Operations Research, Vol 9, pp. 279-286, World Publishing Corporation, Beijing, 2008. (ISTP)
9. **Xiaojian Shao,** Yingjie Tian, Naiyang Deng. SVM-based Automatic Classification for Protein Structural Domain. In Proceedings of 1st International Symposium on Optimization and Systems Biology, Lecture Notes in Operations Research, Vol 7, pp. 341-350, World Publishing Corporation, Beijing, 2007. (ISTP)

在校期间所获奖励 荣获2010-2011年中国农业大学研究生科研成就奖