ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 2,Mar. 2018,pp. 195－206
DOI:10. 16337/j. 1004-9037. 2018. 02. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 识别蛋白质配体绑定残基的生物计算方法综述

於东军　朱一亨　胡　俊

(南京理工大学计算机科学与工程学院,南京,210094)

**摘　要:** 蛋白质与配体相互作用在生命过程中是普遍存在且不可或缺的,这种相互作用在生物分子的识别和信号传递过程中起着非常重要的作用。识别出蛋白质与配体相互作用的绑定残基对蛋白质功能研究、药物设计和筛选都有着重要的科学意义,而生物计算方法是蛋白质与配体绑定残基预测研究中的一种重要手段。本文首先给出了蛋白质与配体相互作用的绑定残基的一般性定义;其次,总结出了一种蛋白质与配体绑定残基预测方法的分类体系,并对其中一些代表性的预测方法进行了简要阐述;再次,给出了蛋白质与配体绑定残基预测研究中常用的数据库和评价指标,并通过在相关数据集上进行实验比较了具有代表性的预测方法的性能;最后,对若干挑战性问题进行分析并预测该领域未来的研究方向,以期对相关研究提供一定的参考。

**关键词:** 蛋白质与配体相互作用;绑定残基预测;分类体系;性能比较

**中图分类号:** TP391　　**文献标志码:** A

## An Overview of Biocomputing Methods of Targeting Protein-Ligand Binding Residues

Yu Dongjun, Zhu Yiheng, Hu Jun

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China)

**Abstract:** Interactions between proteins and ligands are ubiquitous and indispensable in the process of life. These interactions play important roles in the biological molecular recognition and in the process of signal transmission. Identifying the binding residues of the protein-ligand interactions has important scientific significance for protein function research, drug design and screening. Biocomputing method has become an important method for the prediction of protein-ligand binding residues. This paper first describes the common definition of the binding residues of the protein-ligand interactions. Next, a systematic category of protein-ligand binding residue prediction is proposed, and several prediction methods in each category are described. This paper then demonstrates the related protein databases and the evaluation indexes, and experimentally compares and analyzes the performances of some representative prediction methods on the corresponding data sets. Finally, the future research directions of protein-ligand binding residue prediction are proposed, which provide some references for relevant researchers engaged in the field of protein-ligand binding residue prediction.

**Key words:** protein-ligand interaction; binding residues prediction; systematic category; performance comparisons

---

**基金项目:** 国家自然科学基金(61772273,61373062)资助项目。
**收稿日期:** 2017-12-28;**修订日期:** 2018-02-27

## 引 言

蛋白质作为生命活动的物质基础之一，是构成一切细胞组织结构的重要组成成分，参与了生物体内许多方面的重要生命过程，是生命活动的重要承担者。因此，对蛋白质的结构、功能以及相互作用等方面进行深入的分析与理解，可以直接、准确地解释各种生命活动现象，亦有助于阐明相关疾病的发病机理，进而指导相应的药物设计[1]。

蛋白质在生命活动过程中不是孤立存在的，需要和其他生物分子、离子等绑定，进而产生相互作用来完成特定的生物功能，这种相互作用在生命过程中普遍存在并且不可或缺[2]。蛋白质所绑定的生物分子、离子等称为配体，如金属离子、小的有机/无机分子、大分子(如蛋白质)、核酸等。在与配体相互作用时，蛋白质中一些关键的氨基酸残基形成一个类似口袋的形状区域，以完成对特定配体的绑定。这些关键的氨基酸残基称为绑定残基(位点)。从一个蛋白质识别出绑定残基，对于理解蛋白质的功能、药物设计、分析生物分子之间的相互作用、指导相关生化实验具有重要意义。

传统上，蛋白质与配体的绑定残基通过生物学实验来测定，此类方法虽然准确，但存在着诸如耗时、昂贵等问题，远远不能满足后基因组时代蛋白质测序工作飞速发展的要求。据统计，当前已测序的蛋白质中，仅 0.6%左右的蛋白质具有配体绑定残基的生物功能注释。为了弥补生物实验方法的不足，研究基于生物计算的蛋白质与配体绑定残基预测方法，以期提升绑定残基的识别速度与精度，是迫在眉睫的。

## 1 蛋白质-配体绑定残基

在著名国际蛋白质结构预测竞赛(Critical assessment of protein structure prediction，CASP)[3-4]与高品质的半人工标注的蛋白质与配体绑定位点数据库 BioLip[5]中，蛋白质与配体绑定残基的定义如下：蛋白质至少有一个重原子与配体分子中任意一个重原子之间距离不超过这两个原子对应的范德华半径之和0.5 Å。图 1 给出了蛋白质与配体绑定残基的示意图。结合图 1，蛋白质与配体绑定残基的具体描述如下。

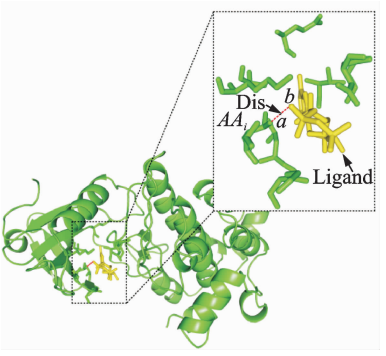


图 1 蛋白质与配体绑定残基示意图

Fig. 1 Schematic diagram of protein-ligand binding residues

假设 $a$ 为蛋白质中第 $i$ 个氨基酸残基 $AA_i$ 中的一个重原子，$b$ 为对应配体(Ligand)中的一个重原子，dis 为重原子 $a$ 与 $b$ 之间的欧式距离，vdw($a$)与 vdw($b$)分别表示重原子 $a$ 与 $b$ 的范德华半径。如果式(1)成立，$AA_i$ 为蛋白质中的一个配体绑定残基；如果 $AA_i$ 中没有任何一个重原子使得式(1)成立，则 $AA_i$ 为非配体绑定残基。

$$\mathrm{dis} \leqslant \mathrm{vdw}(a) + \mathrm{vdw}(b) + 0.5\ \text{Å} \tag{1}$$

蛋白质-配体绑定残基预测问题的目标是从蛋白质信息出发识别出这些关键的氨基酸残基，即配体绑定残基。如何借助已标注配体绑定残基信息的蛋白质来预测待测蛋白质的配体绑定残基、如何应对类不平衡学习问题对配体绑定残基识别精度的影响以及如何利用机器学习算法提升预测精度等问题均是蛋白质-配体绑定残基预测研究中亟需克服的挑战。为了应对上述挑战，相关学者在近几十年来进行了大量的研究工作，并提出了多种多样的预测方法，促进了蛋白质-配体绑定残基预测研究的发展。

## 2 国内外研究现状

随着蛋白质序列数据库、蛋白质结构数据库及相关蛋白质-配体数据库中数据的不断丰富，利用生

物计算方法进行蛋白质-配体绑定残基的预测已经成为后基因组时代生物信息学研究中一个非常活跃的领域。近几十年来,不断涌现出新的生物计算方法来预测蛋白质-配体绑定残基。根据所使用的特征类型,现有的蛋白质-配体绑定残基预测方法大致可分为 3 种类型[6-7]:基于蛋白质结构信息的生物计算方法、基于蛋白质序列信息的生物计算方法以及基于结构与序列信息的混合生物计算方法。

其中,基于蛋白质结构信息的生物计算方法又可进一步地分为基于结构模板匹配的预测方法、基于空间几何的预测方法和基于能量的预测方法;基于蛋白质序列信息的生物计算方法可分为基于序列模板匹配的预测方法、基于序列特征的机器学习的预测方法和序列模板匹配与机器学习的混合方法;基于结构与序列信息的混合生物计算方法可细分为基于结构与序列信息的模板匹配方法、基于结构与序列特征的机器学习方法和基于已有预测方法的整合提升方法。这种识别蛋白质-配体绑定残基的生物计算方法的分类体系如图 2 所示。

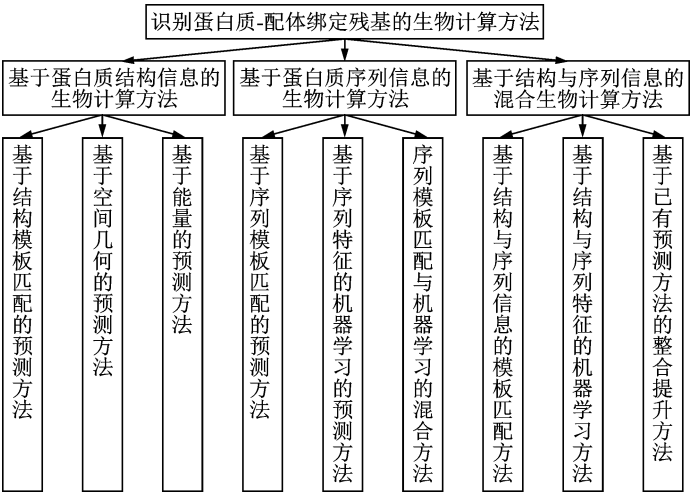


图 2 蛋白质与配体绑定残基计算方法分类图

Fig. 2 Classification of protein-ligand binding residues prediction methods

## 2.1 基于蛋白质结构信息的生物计算方法

在早期阶段,基于蛋白质结构信息的配体绑定残基预测方法占据主导地位。该类方法按计算方法的不同可以分为如下 3 个子类方法。

### 2.1.1 基于结构模板匹配的预测方法

在生物信息学领域中,研究学者普遍认为具有相似结构的蛋白质往往具有相似的生物功能[8]。这也是基于结构模板匹配的预测方法的思想来源。优秀的基于结构模板匹配的预测方法有 3DLigandSite[9]、FINDSITE[10]与 FunFOLD[11]等。为了识别待测蛋白质中的配体绑定残基,基于结构模板匹配的预测方法首先使用蛋白质结构对齐方法(如 Dali[12]、MAMMOTH[13]与 TM-align[14]等)来评价所有已标注配体绑定位点蛋白质与待测蛋白质之间的结构相似性;而后,以结构相似性程度作为依据,对已知配体绑定位点蛋白质进行排序筛选,选择出若干个蛋白质作为模板,同时提取这些模板与待测蛋白质的结构对齐信息;最终,根据这些对齐信息,按照特定的配体绑定位点识别规则,来预测待测蛋白质中潜在的配体绑定残基[9-10, 15]。

以 3DLigandSite[9]为例来说,3DLigandSite 首先使用 MAMMOTH[13]来评价待测蛋白质与所有模板蛋白质之间的结构相似度,并选择相似程度最高的 25 个模板蛋白质以及对应的配体结构信息;再根据选中的模板蛋白质与待测蛋白质的结构对齐信息,将所有模板蛋白质对应的模板配体结构叠加到待

测蛋白质的结构上;然后使用单联动聚类算法(Single linkage clustering,SLC)将所有叠加后的模板配体结构进行聚类,形成多个模板配体簇;最后选择具有最多模板配体的簇,并根据该簇中全部模板配体信息判断待测蛋白质中的每个残基是否为配体绑定残基。根据上面的描述可知,这类预测方法的预测精度完全取决于待测蛋白质与模板蛋白质之间的结构相似程度。

2.1.2 基于空间几何的预测方法

根据图 1 所示的蛋白质-配体绑定残基定义,可知配体绑定残基识别精度与空间位置的关系密切。通过空间几何测量的方法可以识别出待测蛋白质中的配体绑定残基。根据蛋白质数据库(Protein data bank, PDB)[16]中蛋白质与配体复合物的详细研究表明,小分子配体倾向于绑定蛋白质表面的凹性区域,尤其是最大最深的那个空洞。因此,大多数基于空间几何的预测方法都侧重于搜索蛋白质上最大的绑定口袋。

基于空间几何的预测方法主要任务是通过计算蛋白质的结构信息中的某种几何测度,来定位待测蛋白质的配体绑定区域,进而确定蛋白质与配体的绑定位点。然而,如何根据蛋白质的结构信息辨别出蛋白质表面的空洞是一件不易的事情。相关学者经过长时间的研究探索,提出了多种具有创造性的方法,下面介绍几个具有代表性的方法。第一种方法是在待测蛋白质周围放置一些规则的网格点并找到那些没有被蛋白质中的原子占据的(空的)网格点[17-19]。例如,LIGSITE[18]首先放置一些规则的三维网格点覆盖在待测蛋白质结构的周围,再从每个网格点分别沿着 $x/y/z$ 轴和该网格的对角线分别画直线,如果一条直线片段的两边均被目标蛋白质所包含,而中间片段没有被包含,那么该区域就是一个候选的绑定口袋。在蛋白质表面上放置空球体是另一种识别空洞的方法[20-21]。例如,SURFNET[21]为了找到最大的空区域,对于任意两个蛋白质原子,只要没有其他原子介于它们之间,就在它们之间放置一个空球。第三种识别空洞的方法是使用 Delaunay 三角化等技术发现蛋白质表面的空隙[22-23]。例如:CASTp[24]综合使用了计算几何学中的 $\alpha$-shape 理论和三角化技术。首先对目标蛋白质进行 Delaunay 三角化,然后根据邻域三角片的法向量方向就可以预测出潜在的绑定口袋。

2.1.3 基于能量的预测方法

基于能量的预测方法的主要目的是根据能量分布情况找到有利于配体分子结合的蛋白质表面的空洞区域[25-30]。该类方法往往会通过设计探针分子并计算探针与周围蛋白质原子之间的交互能量来识别待测蛋白质中的配体绑定残基[31]。许多基于能量的预测方法同时也是基于网格的,因为它们会将探针放置在蛋白质表面的空网格上以执行对应的能量计算[25, 27, 30]。例如,SITEHOUND[26]使用分子间的交互作用力场计算待测蛋白质与探针之间的相互作用力,具有高能量的网格点被选中用于进一步地聚类分析,并根据聚类结果确定潜在的配体绑定位点。

一般来说,基于能量的预测方法的多样性会比基于空间几何的预测方法少很多。不同的基于能量的预测方法之间的区别主要在于探针的设计方法以及将探针分布在蛋白质表面的方法。探针的复杂度直接影响着预测精度与计算时间复杂度。因此,许多基于能量的预测方法的对应服务网站上提供了多种类型的探针,以应对不同的用户需求。例如,SITEHOUND[26]服务网站(http://scbx.mssm.edu/sitehound/sitehound-web/input.html)允许用户选择 4 种不同类型的探针来计算交互能量;FTSite[32]服务网站(http://ftsite.bu.edu/)提供了 16 种不同的小分子探针来确定合理的网格簇。

## 2.2 基于序列信息的生物计算方法

目前已经测出三维结构的蛋白质数量要远远少于已经测出序列的蛋白质数量。这就意味着还有大量的蛋白质只知道序列信息,而不知道结构信息,从而导致它们的配体绑定位点并不能使用基于结构的方法来进行预测。尽管基于同源建模的工具(如:MODELLER[33]、Rosetta[34]与 I-TASSER[35]等)可以通过蛋白质序列信息预测出对应的结构信息,但是通过建模预测出的蛋白质结构的精度并不能得到很好的保证,且潜在的配体绑定区域的局部结构的精度甚至会更低。另外,还有些蛋白质没有三维结构已知的同源蛋白质,从而导致它们无法使用同源建模工具来预测结构信息。因此,直接从蛋白质序列信息

出发来预测蛋白质与配体绑定位点引起了相关学者的广泛关注。近年涌现出许多基于序列信息的预测方法,如基于序列模板匹配的预测方法、基于序列特征的机器学习的预测方法和序列模板匹配与机器学习的混合方法等。

2.2.1 基于序列模板匹配的预测方法

与前面所述的基于结构模板匹配的预测方法类似,基于序列模板匹配的预测方法主要是根据蛋白质序列与序列之间的同源性或相似性来从对应的数据库中搜索到一个或多个优秀的序列模板,再根据搜索的序列模板对应的配体绑定位点信息来预测蛋白质中潜在的配体绑定位点。例如,文献[36]中描述了一种基于 BLAST 的预测方法,该方法使用 BLAST 序列对齐工具[37]从已标注绑定位点信息的蛋白质数据集合中选择一个序列相似性最高、*E*-value 最低的蛋白质作为模板,并得到该蛋白质模板与待测蛋白质之间的序列对齐信息,最后将待测蛋白质中所有被模板蛋白质中绑定残基对齐上的残基判定为绑定位点。又如,文献[38]中描述的 S-SITE 方法通过比较待测蛋白质的序列谱文件与已标注绑定位点的蛋白质数据库(即 BioLip[5])中的每条蛋白质的序列谱文件之间的相似性来衡量两条对应蛋白质之间的相似性,并按照相似性越高得分越高的方式进行打分;再从数据库中选择与待测蛋白质相似得分超过固定阈值的蛋白质作为模板集;在这个集合中,有超过 25%的蛋白质模板中存在绑定位点对齐于同一个待测蛋白质中的氨基酸残基,该残基则被判定为一个潜在的配体绑定残基。

2.2.2 基于序列特征的机器学习的预测方法

基于序列模板匹配的预测方法对序列的同源性是具有很强的依赖性。当找不到同源性高的模板蛋白质时,预测的配体绑定残基的可信度是不可靠的。为了降低预测方法对于序列同源性的依赖性同时充分挖掘数据中的隐含信息,相关学者试图通过引入机器学习算法来构建预测模型,进而对蛋白质-配体绑定残基进行识别。基于序列特征的机器学习的预测方法的主要步骤如下:首先抽取已标注配体绑定位点的蛋白质序列的有效信息(如氨基酸组成成分信息、蛋白质进化信息与基于序列预测的二级结构信息等),然后根据这些信息构建氨基酸残基的特征向量,生成对应的氨基酸残基样本,进而构建训练样本集合。最后使用恰当的机器学习算法在该训练样本集合上训练预测模型,用于预测任意一个待测蛋白质氨基酸残基是否为配体绑定位点。例如,L1pred 预测方法[39]使用 8 种基于序列信息的特征(包括残基类型、重叠属性、平均累积疏水性、预测的二级结构、预测的接触表面积、Jensen-Shannon 背离得分等)来组成特征向量,作为 L1-logreg 算法[40]的输入信息来预测酶蛋白的催化残基(即底物的绑定残基);ATPint[41]使用蛋白质进化信息、基于统计的氨基酸残基的疏水性、基于统计的氨基酸残基的溶剂可及性面积等信息来构建特征向量,并使用支持向量机(Support vector machine, SVM)算法[42]训练预测模型,用于识别待测蛋白质中与 ATP 相互作用的绑定残基;VitaPred 预测方法[43]使用 PSI-BLAST 工具[37]从蛋白质序列信息中抽取位置特异性得分矩阵作为蛋白质进化信息的特征源来构建氨基酸残基的特征向量,再将训练集中的蛋白质序列转换为以氨基酸为单位的训练样本集合,最后使用 SVM 算法学习预测模型,来辨别待测蛋白质中与维他命相互作用的绑定残基。

2.2.3 基于序列模板匹配与机器学习的混合方法

为了融合基于序列模板匹配的预测方法与基于机器学习的预测方法的优点,相关学者提出了基于序列模板匹配与机器学习的混合方法。此类预测方法将上述两种方法按照特定方式相结合来预测蛋白质与配体相互作用的绑定残基。最具代表性的方法是文献[36]中提出的 NsitePred 预测方法。NsitePred 方法实现了一种基于 SVM 算法的预测方法和一种基于 BLAST 工具的预测方法,并通过求得这两种方法的输出概率的均值来融合它们。2.2.1 节简要描述了基于 BLAST 工具的预测方法的预测步骤。接下来,本文简述 SVMPred 的工作步骤:(1)从蛋白质序列信息中抽取预测的二级结构信息、预测的可接触表面积和二面角信息、位置特异性得分矩阵信息、终端标识信息(即蛋白质序列的首尾 3 个残基为 1,其余残基为 0)、二级结构片段指标、残基保守性得分和重要残基配对信息;(2)使用大小为 17 的滑动窗口技术,抽取训练集中每个蛋白质的每个残基的特征向量,从而组成训练样本集合;(3)使

用 SVM 算法在训练样本集合上学习得到预测模型;(4)对于任意一个待测蛋白质序列,先抽取该蛋白质中每个残基的特征向量,并输入到训练好的预测模型中进行蛋白质与配体绑定残基的预测。

## 2.3 基于结构与序列的混合生物计算方法

如前所述,基于蛋白质结构信息的生物计算方法与基于蛋白质序列信息的生物计算方法都取得了不错的研究成果。为了进一步提升蛋白质与配体绑定残基识别精度,一些研究学者提出了基于蛋白质结构与序列信息的混合生物计算方法。该方法试图将蛋白质结构与序列信息相融合,构建识别精度更高的方法。具体来说,基于结构与序列信息的混合生物计算方法是将从蛋白质结构上抽取的特征信息与从序列上抽取的特征信息相结合,形成一个具有鉴别能力的特征表示,并使用合适的计算方法来构建预测模型,从而判定待测蛋白质中的每个氨基酸残基是否为配体绑定残基。按照构建模型方法的不同,该类方法又可细分为基于结构与序列信息的模板匹配方法、基于结构与序列特征的机器学习方法以及基于已有预测方法的整合提升方法。

### 2.3.1 基于结构与序列信息的模板匹配方法

基于结构与序列信息的模板匹配方法是首先通过联合待测蛋白质与模板蛋白质之间的结构匹配信息与序列匹配信息对每个模板蛋白质进行综合评价并给出该模板的得分,然后选择一个或多个在结构与序列两个层面都与待测蛋白质相似的蛋白质模板,最后根据选中的模板进行蛋白质与配体绑定残基的识别。在此类方法中,最具代表性的是文献[38]中提出的 TM-SITE 方法。

TM-SITE [38] 的具体步骤如下:(1)使用结构对齐工具 TM-align[14] 将待测蛋白质与 BioLip 数据库[5] 中的每条模板蛋白质在结构层面上进行对齐;(2)在上述对齐结果的基础上结合序列层面的进化保守信息和结构层面的空间距离等信息构造出了一个合理、有效的评价函数;(3)利用评价函数对每条蛋白质序列进行相似性评价,并从 BioLip 数据库中选择满足条件的蛋白质结构;(4)根据选中的蛋白质对应的已标注的绑定位点信息进行待测蛋白质与配体绑定位点的预测。

### 2.3.2 基于结构与序列特征的机器学习方法

顾名思义,该方法主要是利用机器学习算法在由结构与序列信息组成的特征空间中构建预测模型,并依此来识别蛋白质与配体绑定残基。此类方法的代表是文献[44]中给出的蛋白质与 DNA 绑定残基预测方法。该方法不仅使用了物化/生化属性、序列保守信息、氨基酸有序/无序信息、预测的二级结构信息等序列层次的特征,还引入了从蛋白质三维结构中计算得到的 5 个结构特征来共同预测蛋白质与 DNA 配体绑定位点。此外,文献[44]为了提升预测精度,使用了最大相关最小冗余方法(Maximum relevance minimum redundancy, mRMR)来进行增量特征选择(Incremental feature selection, IFS)的研究,并确定最终的特征空间。最后,文献[44]将已知 DNA 绑定位点的蛋白质中的氨基酸残基转换成特征空间中的训练样本,并使用 SVM 算法训练预测模型,从而鉴别出待测蛋白质中潜在的、与 DNA 相互作用的绑定残基。

### 2.3.3 基于已有预测方法的整合提升方法

基于已有预测方法的整合提升方法是通过将已有的蛋白质与配体绑定残基预测方法的输出值进行直接融合或当作新的特征再学习一个更深层次的预测模型,以期提升最终的预测精度。具有代表性的方法有:ConCavity[45]、MetaDBSite[46] 和 COACH[38] 等。ConCavity[45] 通过使用 3 个已有的蛋白质与配体绑定残基预测工具(LIGSITE[18]、SURFNET[21] 和 PocketFinder[47])的预测结果与抽取的进化保守信息相结合来共同识别蛋白质与配体的绑定残基;MetaDBSite[46] 主要利用 SVM 算法整合 6 个有效的在线服务器(即:DISIS[48]、DP-Bind[49]、DNABindR[50]、BindN[51]、Bindn-rf[52] 以及 DBS-Pred[53])的输出结果来预测蛋白质与 DNA 相互作用的配体绑定残基;COACH[38] 使用 SVM 算法整合了 S-SITE[38]、TM-SITE[38]、COFACTOR[54]、FINDSITE[10] 以及 ConCavity[45] 5 种预测方法来融合蛋白质结构与序列信息,从而大幅度提升了蛋白质与配体绑定位点的预测精度。

## 3 类不平衡学习问题

由蛋白质与配体绑定残基预测研究问题的国内外研究现状可知,基于机器学习算法(如 SVM 算法)的预测方法在该项科学研究中占据了主导地位。而基于机器学习算法的蛋白质配体绑定残基预测研究是一个典型的二类类不平衡学习问题。如图 3 所示,蛋白质中只有少部分的关键残基会与配体产生相互作用(红色标注出的绑定氨基酸残基),而大多数的残基(绿色标注出的非绑定氨基酸残基)并没有直接地与配体相互接触。就图 3 中的 2xef:A 蛋白质而言,绑定残基的数目为 16,而非绑定残基的数目为 225,非绑定残基数目是绑定残基数目的 14 倍多。

由于大多数的机器学习算法都是基于类平衡的假设来研究设计的,直接使用这些算法来识别蛋白质与配体绑定残基会不可避免地受到类不平衡数据的影响,从而导致最终的预测结果偏向于多数类(即非绑定残基),而少数类(配体绑定残基)的检出率低,并不能达到实际应用的要求。

近年来,为了减轻类不平衡问题在蛋白质与配体绑定残基预测研究中带来的负面影响,相关学者做出了大量的研究工作。其中,最普遍的解决方法是通过样本采样技术对类不平衡数据集中的样本分布进行重新调整来获取类平衡数据,并使用传统机器学习算法进行模型训练。这里的样本采样技术又可大致分为上采样方法(Over-samping method)与下采样方法(Under-sampling method)[55]。上采样方法是通过某种规则根据现有的样本分布情况增加少数类的样本数目来缓解原有的类不平衡程度。与上采样方法相反,下采样方法是使用特定规则根据现有的样本分布减少多数类的样本数目,以期改变原来的类不平衡现象。

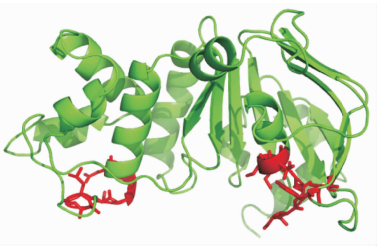

图 3 蛋白质(2xef:A)的绑定残基(红色)与非绑定残基(绿色)

Fig. 3 Binding residues (red) and non-binding residues (green) of protein 2xef:A

使用上采样方法来处理类不平衡问题的代表性方法是 TargetSOS[56]。文献[56]中设计出了一种有监督的上采样方法(Supervised over-sampling method, SOS)。该方法首先使用 SVM 算法在原始的类不平衡数据集上训练一个初始的预测模型;该模型用于评估每一个人工合成的少数类样本的质量并给出相应的得分,再根据该得分判定对应的人工样本是否可以加入到训练样本集合中去;上述这种"合成-评估-判定"的过程一直重复至训练样本集合中的不平衡程度减轻到预期值。基于 SOS 方法,文献[56]中实现了 TargetSOS 方法用于预测蛋白质与 5 种核苷酸的配体绑定残基。

使用下采样方法来处理类不平衡问题的预测方法有:TargetATP[57]、TargetS[7] 和 IonCom[58] 等。以 TargetS 方法[7] 为例,该方法使用随机下采样(Random under-sampling, RUS)方法来缓解类不平衡数据所带来的负面影响。由于下采样方法往往会丢失数据中的有效信息,因此,在 TargetS 中 RUS 被多次使用来获得多个不同的较为平衡的训练样本集合,再使用 SVM 算法在每个平衡训练集上训练对应的子预测模型,最后使用一种改良的 AdaBoost 算法(Modified AdaBoost, MAdaBoost)将多个子模型集成为最终的预测模型。

## 4 蛋白质-配体绑定残基预测实验评测

### 4.1 相关的蛋白质数据库

用于进行蛋白质与配体绑定残基预测研究的数据集几乎都来源于 PDB 数据库[16, 59] 或 BioLip 数据库[60]。

PDB 数据库[16, 59] 是在 1971 年由美国 Brookhaven 国家实验室建立的全球统一的生物大分子(包括蛋白质与配体的复合物)的三维结构信息档案库。PDB 数据库中收集的结构信息主要来源于 X 射线晶体衍射、核磁共振、电子显微镜等结构测定技术。目前,PDB 数据库的更新与维护是由结构生物信息学

研究合作组织(Research collaboratory for structural bioinformatics, RCSB)负责的。RCSB 的主服务器和全球的镜像服务器提供数据库的查询和下载服务。除此之外,PDB 数据库还可以从发行的光盘获得。截止至 2018 年 2 月 5 日,PDB 数据库[60]中含有 137 322 个蛋白质结构数据(数据来源于 http://www.rcsb.org/stats/growth/overall)。使用 Rasmol[61]、PyMol[62]以及 Jmol[63]等软件可以在计算机上按照 PDB 文件内容显示生物大分子的空间结构。

BioLip 数据库[5]是于 2013 年根据 PDB 数据库中的蛋白质结构信息构建的,它是半人工标注的高品质数据库。BioLip 旨在构建最全面的、准确的、服务于蛋白质与配体对接[64]、虚拟筛选[65]以及蛋白质功能注释(包括蛋白质与配体绑定位点的识别)[66]的数据库。该数据库收集了 PDB 数据库中所有与生物学相关的蛋白质与配体绑定位点信息。

### 4.2 评价指标

蛋白质与配体绑定残基预测是一个典型的二类分类问题。因此,用于评价二类分类精度的指标也适用于评估蛋白质与配体绑定残基的预测精度。常见的评价指标有:敏感性(Sensitivity, Sen)、特异性(Specificity, Spe)、精确性(Accuracy, Acc)、查准率(Precision, Pre)和马氏相关系数(Matthew′s correlation coefficients, MCC)。它们的定义如下

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{6}$$

式中:TP (True positives)和 FN (False negatives)分别表示正样本(即绑定位点)被正确预测和被错误预测的数目;而 TN (True negatives)和 FP (False positives)分别表示负样本(即非绑定位点)被正确预测和被错误预测的数目。除了上述 5 个评价指标外,受试者工作特征曲线的面积(Area under the receiver operating characteristic curve, AUC)也常用于评价预测方法的总体性能。

### 4.3 几种常见蛋白质配体绑定残基预测方法的性能比较

考虑到有些预测方法为配体特异性的预测方法(即专门预测某种特定的配体类型的绑定残基,如 ATPint[41]被专门用于预测蛋白质与 ATP 的绑定残基),而有些预测方法为配体普适性的预测方法(即可用于预测蛋白质与所有配体的绑定残基,如 COACH[38]等),本节将配体特异性的预测方法和配体普适性的预测方法分开进行比较。

在配体特异性层次的比较上,以蛋白质与 ATP 绑定残基预测问题为例来比较 ATPint[41]、NsitePred[36]、TargetATP[57]与 TargetSOS[56]的预测性能。

表 1 给出了 ATPint[41]、NsitePred[36]、TargetATP[57]与 TargetSOS[56]在独立测试数据集合 ATP17 上的预测性能。ATP17 数据集是文献[36]给出的独立测试集合,它包含 17 个与 ATP 绑定的蛋白质。从表 1 中,可以看出 TargetSOS 取得了最好的预测性能(Sen=0.54, Spe=0.99, Acc=0.98 和 MCC=0.60)。主要原因是 TargetSOS 中使用 SOS 算法有效地缓解了类不平衡学习问题带来的负面影响。由于 TargetATP 中也使用了 RUS 与 MadaBoost 相结合的方法来减轻类不平衡带来的影响,所以 TargetATP 也取得了较 NsitePred(没有关注类不平衡问题)更好的预测性能。与 NsitePred 相比,TargetSOS 与 TargetATP 在 MCC 指标上分别提高了 25.0%和 12.5%。而作为第一个基于序列信息

的蛋白质与 ATP 绑定残基预测方法,ATPint 仅仅取得了 0.07 的 MCC 值,主要原因在于两个方面:(1)用于训练预测模型的蛋白质数据较少;(2)类不平衡学习问题未得到应有的重视。

在配体普适性层次的比较上,直接引入了文献[38]中的 ConCavity[45]、FINDSITE[10]、COFACTOR[54]、S-SITE[38]、TM-SITE[38]和 COACH[38]比较结果(见表 2)。从表 2 中可以看出整合了多个已有预测方法的 COACH 取得了最优的预测性能,它的 MCC 值(0.54)相比于表现第二的 TM-SITE 提升了 12.5%。值得注意的是,S-SITE 虽然是一个基于序列信息的预测方法,但也取得了较好的预测性能(Sen=0.58, Pre=0.45 以及 MCC=0.45),这说明序列中也具有很好的鉴别信息。

**表 1 ATPint、NsitePred、TargetS 与 TargetSOS 在 ATP17 数据集[36]上的独立验证实验结果比较**

**Tab. 1 Performance comparisons of ATPint, NsitePred, TargetS, and TargetSOS on the independent testing dataset ATP17[36]**

| 预测方法 | Sen | Spe | Acc | MCC |
|---|---|---|---|---|
| ATPint* | 0.51 | 0.66 | 0.66 | 0.07 |
| NsitePred* | 0.46 | 0.99 | 0.97 | 0.48 |
| TargetATP# | 0.49 | 0.99 | 0.97 | 0.54 |
| TargetSOS+ | 0.54 | 0.99 | 0.98 | 0.60 |

* 数据摘引自文献[36];# 数据摘引自文献[57];+ 数据摘引自文献[56]。

**表 2 已有预测方法在 COACH500[38]上的独立验证实验结果比较**

**Tab. 2 Performance comparisons of existing predictors on the independent testing dataset COACH500 [38]**

| 预测方法 | Sen | Pre | MCC |
|---|---|---|---|
| ConCavity* | 0.51 | 0.23 | 0.26 |
| FINDSITE* | 0.49 | 0.44 | 0.42 |
| COFACTOR* | 0.39 | 0.56 | 0.42 |
| S-SITE* | 0.58 | 0.45 | 0.45 |
| TM-SITE* | 0.49 | 0.57 | 0.48 |
| COACH* | 0.63 | 0.54 | 0.54 |

* 数据摘引自文献[38]。

## 5 总结与展望

准确地识别出蛋白质中的配体绑定残基,对于理解生命活动中某些未知蛋白质的功能和生命活动现象的本质、进行疾病的诊断以及病理研究具有重要的实际价值。例如,在蛋白质功能预测方面,如果识别出蛋白质配体绑定残基以及相互作用形成的绑定口袋形状,就可以推断出该蛋白质的具体生物功能;在疾病的诊断和药物研发方面,蛋白质与配体相互作用的绑定残基通常也是药物的靶点,准确识别配体绑定残基对于病理分析以及药物研发都具有重要的作用。

由于蛋白质与配体绑定残基预测研究在蛋白质功能及医药研究中具有潜在的应用价值,所以利用生物实验来识别配体绑定残基的方法一直是过去系统生物学中的一项主要内容。然而,伴随蛋白质数据的与日剧增,生物实验方法已经难以满足后基因时代人类对生物体行为理解、蛋白质功能预测和药物设计的需要。因此,生物计算方法无疑是解决该问题的有效途经,它能够及时、高效地进行配体绑定残基识别。尽管近年来人们已经将机器学习、数据挖掘中的一些理论和方法运用于蛋白质与配体绑定残基预测研究中,并且该项预测研究也取得了长足进步,但是所取得的蛋白质与配体绑定位点预测精度与实际应用之间还有很长的距离,该预测问题仍然是生物信息学中一个具有挑战性的科学难题。为了进一步地提升预测性能,以下几个方面研究方向是潜在的突破口。

(1) 从当前已有的成果来看,来自于蛋白质的具有鉴别能力的特征比较有限,这就使得蛋白质与配体绑定位点预测问题的研究陷入了特征不足的瓶颈。因此,从蛋白质序列和结构中抽取更具鉴别性的特征信息,是突破性地提升蛋白质与配体绑定位点预测性能的关键步骤。如何抽取更具鉴别性信息的蛋白质特征是主要研究方向之一。

(2) 位于同条蛋白质的不同氨基酸残基之间的相关性并没有得到充分关注。当前已有的方法一般是在训练阶段使用滑动窗口将相邻的氨基酸残基的信息结合到待判定残基的特征向量中,而在预测阶段时,同一蛋白质中的任意两个残基都是独立地进行预测。这就使得在预测过程大量的隐含在同一蛋

白质中不同氨基酸之间的关联信息被丢失。如何抽取同一蛋白质的不同氨基酸之间的关联信息,并应用于蛋白质与配体绑定位点预测问题,以期提升预测性能,是研究的主要方向之一。

(3) 随着蛋白质数据的日积月累,海量蛋白质数据的时代已经到来。在蛋白质大数据时代背景下,如何利用海量蛋白质数据进行深度学习,充分挖掘隐含在数据中的有用信息,构建具有深度学习能力的蛋白质与配体绑定残基预测模型,是大幅度提升预测性能的重要研究方向。所以,借鉴深度学习算法在图像处理、视频跟踪以及推荐系统等领域中取得的成果,并提出适合于蛋白质与配体绑定位点预测的深度学习模型是未来研究的主要方向之一。

**参考文献:**


[1] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool [J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[2] Chen K, Mizianty M J, Kurgan L. ATPsite: Sequence-based prediction of ATP-binding residues [J]. Proteome Science, 2011, 9(1): S4.

[3] Schmidt T, Haas J, Cassarino T G, et al. Assessment of ligand-binding residue predictions in CASP9 [J]. Proteins: Structure, Function, and Bioinformatics, 2011, 79(S10): 126-136.

[4] Gallo C T, Bordoli L, Schwede T. Assessment of ligand binding site predictions in CASP10 [J]. Proteins: Structure, Function, and Bioinformatics, 2014, 82(S2): 154-163.

[5] Yang J, Roy A, Zhang Y. BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions [J]. Nucleic Acids Research, 2013, 41(D1): D1096-D1103.

[6] Liu R, Hu J. HemeBIND: A novel method for heme binding residue prediction by combining structural and sequence information [J]. BMC Bioinformatics, 2011, 12(1): 207.

[7] Yu D J, Hu J, Yang J, et al. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering [J]. IEEE ACM T Comput Bi, 2013, 10(4): 994-1008.

[8] Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure [J]. Nature Reviews Molecular Cell Biology, 2007, 8(12): 995.

[9] Wass M N, Kelley L A, Sternberg M J. 3DLigandSite: Predicting ligand-binding sites using similar structures [J]. Nucleic Acids Research, 2010: gkq406.

[10] Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation [J]. Proceedings of the National Academy of Sciences, 2008, 105(1): 129-134.

[11] Roche D B, Tetchner S J, Mcguffin L J. FunFOLD: An improved automated method for the prediction of ligand binding residues using 3D models of proteins [J]. BMC Bioinformatics, 2011, 12(1): 160.

[12] Holm L, Sander C. Dali: A network tool for protein structure comparison [J]. Trends in Biochemical Sciences, 1995, 20(11): 478-480.

[13] Ortiz A R, Strauss C E, Olmea O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison [J]. Protein Science, 2002, 11(11): 2606-2621.

[14] Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score [J]. Nucleic Acids Research, 2005, 33(7): 2302-2309.

[15] Roy A, Zhang Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement [J]. Structure, 2012, 20(6): 987-997.

[16] Rose P W, Prlic′ A, Bi C, et al. The RCSB protein data bank: Views of structural biology for basic and applied research and education [J]. Nucleic Acids Research, 2015, 43(D1): D345-D356.

[17] Huang B, Schroeder M. LIGSITE csc: Predicting ligand binding sites using the connolly surface and degree of conservation [J]. BMC Structural Biology, 2006, 6(1): 19.

[18] Hendlich M, Rippmann F, Barnickel G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins [J]. Journal of Molecular Graphics and Modelling, 1997, 15(6): 359-363.

[19] Weisel M, Proschak E, Schneider G. PocketPicker: Analysis of ligand binding-sites with shape descriptors [J]. Chemistry Central Journal, 2007, 1(1): 7.

[20] Brady G P, Stouten P F. Fast prediction and visualization of protein binding pockets with PASS [J]. Journal of Computer Aided Molecular Design, 2000, 14(4): 383-401.

[21] Laskowski R A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions [J].

Journal of Molecular Graphics, 1995, 13(5): 323-330.
[22] Le G V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection [J]. BMC Bioinformatics, 2009, 10(1): 168.
[23] Zhu H, Pisabarro M T. MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets [J]. Bioinformatics, 2010, 27(3): 351-358.
[24] Dundas J, Ouyang Z, Tseng J, et al. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues [J]. Nucleic Acids Research, 2006, 34(Suppl 2): W116-W118.
[25] Ghersi D, Sanchez R. EasyMIFS and SiteHound: A toolkit for the identification of ligand-binding sites in protein structures [J]. Bioinformatics, 2009, 25(23): 3185-3186.
[26] Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: A server for ligand binding site identification in protein structures [J]. Nucleic Acids Research, 2009, 37(Suppl 2): W413-W416.
[27] Silberstein M, Dennis S, Brown L, et al. Identification of substrate binding sites in enzymes by computational solvent mapping [J]. Journal of Molecular Biology, 2003, 332(5): 1095-1113.
[28] Laurie A T, Jackson R M. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites [J]. Bioinformatics, 2005, 21(9): 1908-1916.
[29] Morita M, Nakamura S, Shimizu K. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures [J]. Proteins: Structure, Function, and Bioinformatics, 2008, 73(2): 468-479.
[30] Ngan C H, Hall D R, Zerbe B, et al. FTSite: High accuracy detection of ligand binding sites on unbound protein structures [J]. Bioinformatics, 2011, 28(2): 286-287.
[31] Xie Z R, Hwang M J. Methods for predicting protein-ligand binding sites [J]. Molecular Modeling of Proteins, 2015: 383-398.
[32] Ngan C H, Hall D R, Zerbe B, et al. FTSite: High accuracy detection of ligand binding sites on unbound protein structures [J]. Bioinformatics, 2012, 28(2): 286-287.
[33] Eswar N, Webb B, Marti-renom M A, et al. Comparative protein structure modeling using modeller [J]. Current Protocols in Bioinformatics, 2006: 5.6. 1-5.6. 30.
[34] Ovchinnikov S, Kinch L, Park H, et al. Large-scale determination of previously unsolved protein structures using evolutionary information [J]. Elife, 2015, 4: e09248.
[35] Yang J, Yan R, Roy A, et al. The I-TASSER Suite: Protein structure and function prediction [J]. Nat Methods, 2015, 12 (1): 7-8.
[36] Chen K, Mizianty M J, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors [J]. Bioinformatics, 2012, 28(3): 331-341.
[37] Altschul S F, Madden T L, Schffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.
[38] Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment [J]. Bioinformatics, 2013, 29(20): 2588-2595.
[39] Dou Y, Wang J, Yang J, et al. L1pred: A sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier [J]. PloS one, 2012, 7(4): e35666.
[40] Koh K, Kim S J, Boyd S. An interior-point method for large-scale l1-regularized logistic regression [J]. Journal of Machine Learning Research, 2007, 8(Jul): 1519-1555.
[41] Chauhan J S, Mishra N K, Raghava G P. Identification of ATP binding residues of a protein from its primary sequence [J]. BMC Bioinformatics, 2009, 10: 434.
[42] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.
[43] Panwar B, Gupta S, Raghava G P. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information [J]. BMC Bioinformatics, 2013, 14(1): 44.
[44] Li B Q, Feng K Y, Ding J, et al. Predicting DNA-binding sites of proteins based on sequential and 3D structural information [J]. Molecular Genetics and Genomics, 2014, 289(3): 489-499.
[45] Capra J A, Laskowski R A, Thornton J M, et al. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure [J]. Plos Comput Biol, 2009, 5(12): e1000585.
[46] Si J, Zhang Z, Lin B, et al. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction [J]. BMC Systems Biology, 2011, 5(Suppl 1): S7.

[47] An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes [J]. Molecular & Cellular Proteomics, 2005, 4(6): 752-761.

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

[49] Hwang S, Gou Z, Kuznetsov I B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins [J]. Bioinformatics, 2007, 23(5): 634-636.

[50] Yan C, Terribilini M, Wu F, et al. Predicting DNA-binding sites of proteins from amino acid sequence [J]. BMC Bioinformatics, 2006, 7(1): 262.

[51] Wang L, Brown S J. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences [J]. Nucleic Acids Research, 2006, 34(Suppl 2): W243-W248.

[52] Wang L, Yang M Q, Yang J Y. Prediction of DNA-binding residues from protein sequence information using random forests [J]. Bmc Genomics, 2009, 10(Suppl 1): S1.

[53] 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 [J]. Bioinformatics, 2004, 20(4): 477-486.

[54] Roy A, Yang J, Zhang Y. COFACTOR: An accurate comparative algorithm for structure-based protein function annotation [J]. Nucleic Acids Research, 2012: gks372.

[55] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[56] Hu J, He X, Yu D J, et al. A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction [J]. PloS One, 2014, 9(9): e107676.

[57] Yu D J, Hu J, Tang Z M, et al. Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling [J]. Neurocomputing, 2013, 104: 180-190.

[58] Hu X, Dong Q, Yang J, et al. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals [J]. Bioinformatics, 2016, 32(21): 3260-3269.

[59] Berman H M, Westbrook J, Feng Z, et al. The protein data bank [J]. Nucleic Acids Res, 2000, 28(1): 235-242.

[60] Rose P W, Prlic′ A, Altunkaya A, et al. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information [J]. Nucleic Acids Research, 2017, 45(D1): D271-D281.

[61] Pikora M, Gieldon A. RASMOL AB-New functionalities in the program for structure analysis[J]. Acta Biochimica Polonica, 2015, 62(3): 629-631.

[62] Delano W L. Pymol: An open-source molecular graphics tolls[J]. CCP4 Newsletter on protein Crystallography, 2002, 40: 82-92.

[63] Herraez A. Biomolecules in the computer: Jmol to the rescue[J]. Biochemistry and Molecular Biology Education, 2006, 34(4):255-261.

[64] Slynko I, Rognan D, Kellenberger E. Protein-ligand docking [M]. [S. l.]: Tutorials in Chemoinformatics, 2017: 355.

[65] Hirata S, Shizu K. Organic light-emitting diodes: High-throughput virtual screening [J]. Nature Materials, 2016, 15: 1056-1057.

[66] Das S, Orengo C A. Protein function annotation using protein domain family resources [J]. Methods, 2016, 93: 24-34.

**作者简介：**

**於东军**(1975-)，男，博士，教授，博士生导师，研究方向：模式识别与智能信息处理、生物信息学，E-mail：njyudj@ njust. edu. cn。

**朱一亨**(1993-)，男，博士研究生，研究方向：生物信息学。

**胡俊**(1989-)，通信作者，男，博士研究生，研究方向：生物信息学。

(编辑：夏道家)