DOI:10.3969/j.issn.1005-202X.2022.11.017

医学生物信息

基于序列的蛋白质-GDP结合位点预测

徐淑坦[1,2],王俊豪[1,2],陈明[1,2]
1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306

【摘要】蛋白质-GDP(Guanosine Diphosphate)结合位点的预测对蛋白质功能注释与新药发现有非常重要作用。为了提高预测蛋白质-GDP结合位点的准确度,提出一种基于序列的蛋白质-GDP结合位点预测方法,使用位置特异性迭代算法进行多序列对比得到位置特异性得分矩阵,通过镜像残基可变滑动窗口方法选取蛋白质序列中每个残基的特征向量,利用CNMW(Clustering NearMiss-2 Weighted)下采样解决数据集正负样本的不平衡问题,最后使用支持向量机进行预测。实验结果显示与传统方法相比,本文方法在马修斯相关系数上有显著提升,表明本文方法的有效性和可行性。

【关键词】蛋白质-GDP结合位点;位置特异性得分矩阵;下采样;滑动窗口;支持向量机
【中图分类号】R318;Q811.4 【文献标志码】A 【文章编号】1005-202X(2022)11-1425-06

Sequence-based prediction of protein-GDP binding site

XU Shutan[1,2], WANG Junhao[1,2], CHEN Ming[1,2]
1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Shanghai 201306, China

Abstract: The prediction of protein-GDP (Guanosine Diphosphate) binding site is significant for protein function annotation and new drug discovery. A sequence-based protein-GDP binding site prediction method is proposed for improving the accuracy of protein-GDP binding site prediction. The method uses a position-specific iterative algorithm for multiple sequence comparison to obtain a position-specific scoring matrix, selects the feature vector of each residue in the protein sequence through the mirror residue-based variable sliding window, solves the imbalance problem of the positive and negative samples of the data set using CNMW (Clustering NearMiss-2 Weighted) under-sampling, and finally realizes the prediction *via* support vector machine. The experimental results showed that compared with traditional methods, the proposed method has a significantly higher Matthews correlation coefficient, indicating its effectiveness and feasibility.

Keywords: protein-GDP binding site; position-specific scoring matrix; under-sampling; sliding window; support vector machine

前 言

二磷酸鸟苷(Guanosine Diphosphate, GDP)是核苷酸的一种,参与了生物中大部分生物化学反应,在DNA复制与转录、跨膜运输、肌肉收缩以及多种代谢过程中都发挥着不可替代的作用。在大多数生物细胞活动中,都需要蛋白质与GDP互相结合来发挥其作用。蛋白质与GDP分子间相互作用通过蛋白质中特定位置的氨基酸残基发生,这些特定的关键氨基酸残基称为配体结合位点。配体结合位点在分子对接、药物靶相互作用、化合物设计、配体亲和力预测,分子动力学等领域都有重要作用[1-5]。因此,蛋白质-GDP结合位点的识别不仅有助于探索分子间相互作用的机制,而且有助于有效地解释疾病的发病机制,为药物的发现和设计提供帮助[6]。

传统的研究通常是使用生物学实验预测蛋白质-GDP结合位点[7-8]。实验往往成本高、耗时、难以推广使用。因此许多研究人员转向了基于计算的蛋白质-配体结合位点预测研究。由于蛋白质-GDP相互作用的重要性以及实验方法的困难性,近年来开发了高效准确的计算方法。计算方法利用机器学习技术和

【收稿日期】2022-05-10
【基金项目】上海海洋大学科研专项(A2-2006-21-200327;A2-2006-21-200208)
【作者简介】徐淑坦,博士后,副教授,研究方向:机器学习、生物信息,E-mail: stxu@shou.edu.cn;王俊豪,硕士在读,研究方向:机器学习、生物信息,E-mail: w18621978045@163.com(徐淑坦和王俊豪共同为第一作者)
【通信作者】陈明,博士,教授,研究方向:机器学习、生物信息、区块链技术,E-mail: mchen@shou.edu.cn

生物信息来预测蛋白质序列中GDP结合残基的位置。根据所涉及的生物信息类别，计算预测方法可以分为3类：基于已知蛋白质结构的分子对接方法[9-10]；基于蛋白质序列信息的预测方法[11-13]；基于序列信息和结构信息的混合方法[14-15]。Hernandez等[16]利用蛋白质与探针之间的相互作用力来预测配体结合位点，使用基于能量的方法识别与配体相互作用的高能量区域，采用不同的探针来表示蛋白质结构，不仅可以识别不同类型的结合位点，还可以对其相互作用特性进行特异性描述。但目前大多数蛋白质的结构信息未知，这导致基于结构的结合位点预测方法并不能普遍使用。

在基于蛋白质序列信息预测结合位点的研究领域有许多成果。Chuanhan等[17]提出基于蛋白质序列的ATP（Adenosine Triphosphate）结合位点的预测方法ATPint，ATPint结合支持向量机（Support Vector Machine, SVM）和位置特异性得分矩阵（Position Specific Scoring Matrix, PSSM），在预测ATP结合位点方面取得了开创性成果。Chen等[18-19]开发了ATPsite和NsitePred，将更多的序列信息添加进来，如氨基酸溶剂可及性、蛋白质二级结构以及残基保守性，进一步提高了模型预测性能。Yu等[20]应用AdaBoost算法解决了数据集不平衡问题，取得了很好的结果。同时，Yu等[21]提出了另一种ATP结合位点预测方法，称为TargetATPsite，该方法结合了分类器集成和残基进化图像稀疏表示。Fang等[22]从PSSM矩阵中提取特征，使用SVM作为分类器来预测ATP结合位点，预测结果的AUC达到0.899，这表明蛋白质序列信息中隐藏着重要的结合特性。

尽管先前的研究取得了重大进展，但仍有改进的空间。蛋白质序列中GDP结合位点的预测是不平衡二分类问题，因为蛋白质序列中GDP非结合残基的数量远多于结合残基的数量。因此直接使用传统的机器学习技术可能导致实验结果较差。为解决这一问题，本研究结合K-means聚类、NearMiss-2算法和加权下采样在处理不平衡数据方面的优点，提出一种基于CNMW（Clustering NearMiss-2 Weighted）下采样的预测方法。首先从蛋白质序列信息入手，提取基于氨基酸残基进化的特征信息，然后通过改进的滑动窗口方法选取以每个残基为中心的邻域残基的特征组合，利用CNMW下采样方法使样本集中的多数类样本和少数类样本数目达到平衡，最后使用5重交叉验证法对SVM进行评估，并使用测试集验证模型的性能，实验结果表明本文方法能有效地提高蛋白质-GDP结合位点预测性能。

1 材料与方法

1.1 数据集

本实验使用3个蛋白质-GDP结合位点标准数据集（表1）。数据集GDP105是Chen等[19]收集的（http://biomine.ece.ualberta.ca/nSITEpred），该数据集由105条非冗余蛋白质序列和105条标签组成，包含1 577个GDP结合位点和36 561个非结合位点。数据集GDP82由82条非冗余蛋白质序列和82条标签组成，包含1 101个GDP结合位点和26 244个非结合位点。数据集GDP14由14条非冗余蛋白质序列组成，包含194个GDP结合位点和4 180个非结合位点。数据集GDP82和GDP14从数据库BioLip获得[23]。将数据集GDP105、GDP82作为训练集，将GDP14作为测试集。

表1 训练集和独立测试集的构成

Table 1 Composition of training set and independent test set

数据集	序列数量	正样本	负样本	负正比值
GDP105	105	1 577	36 561	23.18
GDP82	82	1 101	26 244	23.84
GDP14	14	194	4 180	21.55

1.2 特征提取

1.2.1 基于氨基酸残基进化的特征提取与标准化 本研究将蛋白质序列的进化信息融入到特征向量中，基于蛋白质序列信息预测蛋白质-GDP结合位点。针对长度为L个氨基酸的蛋白质序列，通过对Swiss-Prot数据库使用位置特异性迭代搜索（Position Specific Iterative Basic Local Alignment Search Tool, PSI-BLAST）算法[24]（E值为0.001，迭代次数为3）对查询序列进行多序列对比，生成PSSM文件，取第3列至第22列构成PSSM矩阵，因为生物体中主要有20种基本的氨基酸，而第3列至第22列分别代表蛋白质序列中氨基酸残基进化为20种氨基酸中其中一种的概率。

长度为L个氨基酸的蛋白质序列P的PSSM矩阵包含L行和20列，表示如下：

$$\boldsymbol{P}_{\mathrm{PSSM}}=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & \cdots & p_{i,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ p_{l,1} & p_{l,2} & \cdots & p_{l,20} \end{bmatrix} \tag{1}$$

其中，$p_{i,j}$代表蛋白质序列P的第i个位置上的氨基酸残基进化为第j种氨基酸的概率。

获得PSSM矩阵后，对所得到的PSSM矩阵中的每个值，使用Logistic函数将其归一化为0~1的范围，Logistic函数定义如下：

$$f(x)=\frac{1}{1+e^{-x}} \tag{2}$$

其中，x是PSSM矩阵中原始值。

1.2.2 基于镜像残基的可变滑动窗口　在判断蛋白质序列中氨基酸残基是否为GDP结合位点时，可以结合该残基的邻域残基的特征。一个残基的特征向量是以它为中心的邻域残基的特征组合。因此，在提取一个氨基酸残基对应的PSSM特征向量时需要采用滑动窗口方法。不同大小的滑动窗口对预测模型的性能是有影响的。滑动窗口越大，意味着提取到的氨基酸残基特征向量包含的邻域特征信息越多，预测模型会得到更好的预测结果，但由于滑动窗口过大，特征向量包含的噪声信息多造成过拟合，导致预测效果降低；滑动窗口越小，意味着特征向量包含的邻域特征信息越少，必然会导致预测模型性能降低。因此，选择合适的滑动窗口大小是影响预测结果的关键问题。

传统的滑动窗口方法在提取蛋白质序列首末两端的部分氨基酸残基特征向量时，在首端和末端分别补上$(L-1)/2$个假残基“X”（L为滑动窗口的大小），这种固定滑动窗口大小的方法会导致氨基酸残基特征向量所包含的邻域信息不足[17]。为了解决这个问题，本研究提出了可变滑动窗口构建氨基酸特征向量。假设R_i是氨基酸序列中的一个残基，R_i的特征向量为一个长度为L的片段S_{R_i}：

$$S_{R_i}=\left\{R_{i-\frac{L-1}{2}},\cdots,R_{i-1},R_i,R_{i+1},\cdots,R_{i+\frac{L-1}{2}}\right\} \tag{3}$$

其中，残基R_i位于相邻残基的中心，其前面有$(L-1)/2$个残基，后面有$(L-1)/2$个残基，这些残基构成一个长度为L的片段。

对于位于蛋白质序列两端的残基，通过镜像可用残基来增加片段的缺失侧。图1显示了缺失残基（用*标记）的特征向量创建方法。其中，$S_{R_1}=\{R^*_{1+\frac{L-1}{2}},\cdots,R^*_3,R^*_2,R_1,R_2,R_3,\cdots,R_{1+\frac{L-1}{2}}\}$，$R^*_{1+\frac{L-1}{2}},\cdots,R^*_3,R^*_2$分别是$R_{1+\frac{L-1}{2}},\cdots,R_3,R_2$的镜像残基；$S_{R_2}=\{R^*_{2+\frac{L-1}{2}-2},\cdots,R^*_2,R_1,R_2,\cdots,R_{2+\frac{L-1}{2}-2},R_{2+\frac{L-1}{2}-1},R_{2+\frac{L-1}{2}}\}$，$R^*_{2+\frac{L-1}{2}-2},\ldots,R^*_2$分别是$R_{2+\frac{L-1}{2}-2},\ldots,R_2$的镜像残基；$S_{R_{\frac{L-1}{2}}}=\{R^*_2,R_1,R_2,\cdots,R_{\frac{L-1}{2}},R_{\frac{L-1}{2}+1},\cdots,R_{\frac{L-1}{2}+\frac{L-1}{2}-1},R_{\frac{L-1}{2}+\frac{L-1}{2}}\}$，$R^*_2$是$R_2$的镜像残基。

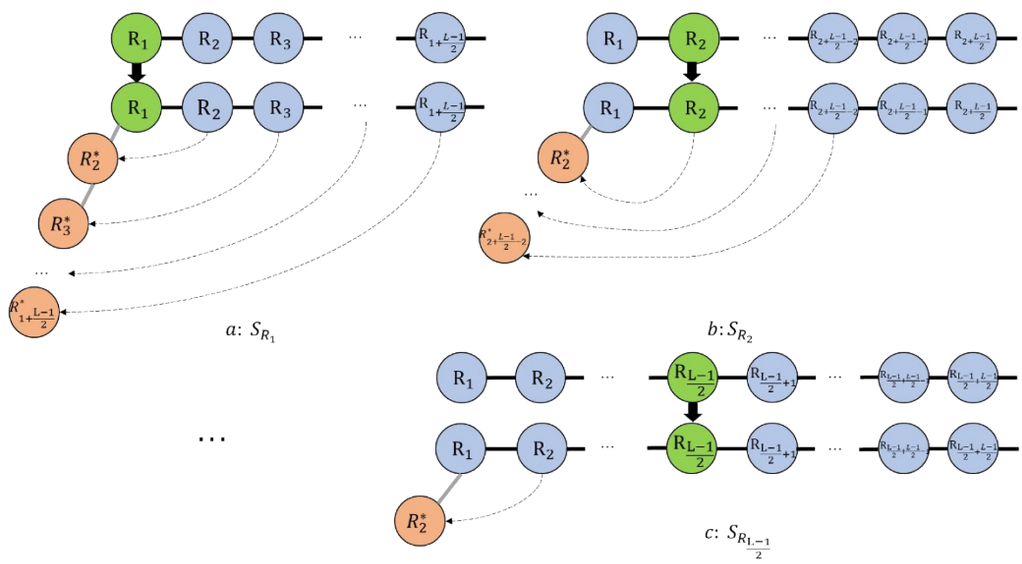

图1　创建缺失残基的过程

Figure 1 Process of creating missing residues

不同大小的滑动窗口所包含的氨基酸数目是不同的，因此输入到训练模型的特征向量信息也是不同的。为了对比不同大小的滑动窗口对模型性能的影响，本研究将滑动窗口大小设置在区间[7, 22]，然后根据模型所得实验结果获得最佳滑动窗口。

1.3 方法

图2显示了本研究提出的蛋白质-GDP结合位点预测的流程。对于一条蛋白质序列，首先使用位置特异性迭代算法得到其PSSM特征矩阵，其次通过基于镜像残基的可变滑动窗口方法提取PSSM特征向

量，然后采用CNMW下采样方法使多数类样本和少数类样本达到平衡，最后使用5重交叉验证法对SVM模型进行评估和预测。

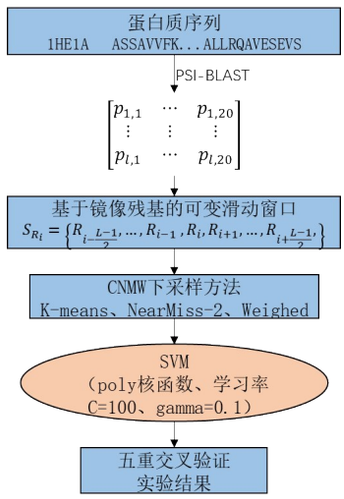

图2 方法流程

Figure 2 Flowchart of the proposed method

1.3.1 CNMW下采样算法 K-means是典型的基于欧式距离的聚类算法，采用距离作为数据对象间相似性度量的标准，即两个数据对象间的距离越小，其相似性越高，将相似性高的的数据对象划分到同一个类簇。每一个簇中的数据对象都可以由K-means算法聚类得出的每个簇中心点来描述。本研究对多数类样本进行K-means算法聚类得到K个类簇，按照NearMiss-2距离[25]为各个簇赋予相应的权重，即多数类样本所具有的第一个权重w_1。然后从样本全局考虑，利用最近邻的思想对样本集中的每一个样本赋予权重，即样本所具有的第二个权重w_2，此时样本集中每个多数类样本有两个不同的权重，把每个多数类样本所对应的两个权重相乘，按照相乘后的权值大小选择多数类样本。

CNMW算法的步骤如下。记总的样本集为P，样本数目为Q，少数类样本集为S，样本数目为M，多数类样本集为T，样本数目为N。根据多数类样本数目N确定K值：

$$K = \left[(2N + 0.25)^{1/2} - 0.5 \right] \tag{4}$$

对多数类样本进行K-means算法聚类得到K个簇及其中心点，根据NearMiss-2距离为K个簇中心点赋予权重。

设样本集$P = \{x_i, y_i\}_{i=1}^{Q}$，$x_i$表示样本集中的一个样本，$y_i$表示对应的类标签，计算每个样本最近邻的$k$个样本，并为其赋予相应的得分，如下所示：

$$\text{score}_{ij} = \begin{cases} 1, x_j 是 x_i 的 k 近邻样本且 y_i = y_j \\ -1, x_j 是 x_i 的 k 近邻样本且 y_i \neq y_j \\ 0, 其他 \end{cases} \tag{5}$$

score_{ij}构成一个$Q \times Q$的得分矩阵，按照公式(6)，可以计算样本集中每个样本的权值：

$$\text{weight}_i = \sum_{j=1}^{n} (\text{score}_{ij} + \text{score}_{ji}) \tag{6}$$

样本集中每个多数类样本有两个权重，将每个多数类样本对应的两个权重相乘得到一个新的权重，将新的权重从大到小排序，并按此顺序选择和少数类样本一样多的多数类样本，并与少数类样本构成新的数据集。

1.3.2 SVM和交叉验证 本实验采用SVM作为分类器解决小样本情况下的二分类问题[26]，其中核函数采用多项式(POLY)核函数，并利用网格搜索法发现学习率(C)为100，多项式核的gamma参数为0.1时，实验能得到较好的结果。

交叉验证法可以从小样本数据集中获得更多的有效信息，而且可以避免过拟合，可用来评估非平衡数据分类模型的预测性能。因此，本研究基于交叉验证法将训练集随机分为5个不相交的子集，选择4个子集作为训练集，剩余一个子集作为测试集，重复此过程5次，直到每个子集都做过测试集为止。将5次预测的平均预测性能作为模型的最终性能。

2 结果与分析

2.1 评价标准

蛋白质-GDP结合位点的预测是二分类问题，所有样本可分为结合位点(正类)和非结合位点(负类)。为了有效地评估该模型的性能，根据模型预测结果的4种情况，可以构建一个混淆矩阵，如表2所示。

表2 混淆矩阵

Table 2 Confusion matrix

预测类别	实际类别	
	实际正类	实际负类
预测正类	TP	FP
预测负类	FN	TN

其中TP(True Positive)是指实际结合残基被正确预测为结合位点的的数量，FP(False Positive)是指实际非结合残基被错误预测为结合位点的的数量，FN(False Negative)是指非结合位点被错误地预测为实际结合残基的数量，TN(True Negative)是指实际非结合残基被正确预测为非结合位点的的数量。在本研究中，采用4个常规的评估标准来检验所提方法的整体性能：准确性(Accuracy, Acc)、灵敏性(Sensitivity, Sen)、特异性(Specificity, Spe)、马修斯相

关系数(Matthews Correlation Coefficient, MCC)。这些评价标准常用于生物信息学研究以评估分类性能。这些标准的定义如下:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{7}$$

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{9}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \tag{10}$$

Acc是对整体准确率的评估;Sen是所有的结合位点被正确识别的比例,衡量了预测模型对结合位点的识别能力;Spe表示的是所有非结合位点中被正确识别的比例,衡量了预测模型正确识别非结合位点的能力;MCC指标考虑混淆矩阵中的真阳性、假阳性、假阴性和真阴性,是衡量二分类模型性能的一个均衡指标。

2.2 滑动窗口选择

滑动窗口的大小会对预测模型的性能产生影响。为了测试不同滑动窗口大小对模型的影响,本实验将滑动窗口大小区间设置为[7, 22],其中只取奇数(体现出中心残基的思想)。图3和图4为滑动窗口大小取不同值时,模型在GDP105数据集和GDP82数据集的预测结果。可以看出,在GDP105数据集中,当滑动窗口大小为7时,预测模型的性能较差,Sen、Spe、Acc和MCC指标分别为66.96%、99.66%、98.25%、0.767,随着滑动窗口逐渐增大,预测模型的性能逐渐升高,当滑动窗口大小为15时,Sen、Acc和MCC指标是最高的,分别为67.72%、98.55%、0.806,只有Spe指标(99.92%)略低,预测模型的整体性能最好,然而当滑动窗口继续增大时,预测模型的性能开始平缓下降。基于以上结果,本实验选择的滑动窗口大小为15,即对应的特征维数为300(15×20)。

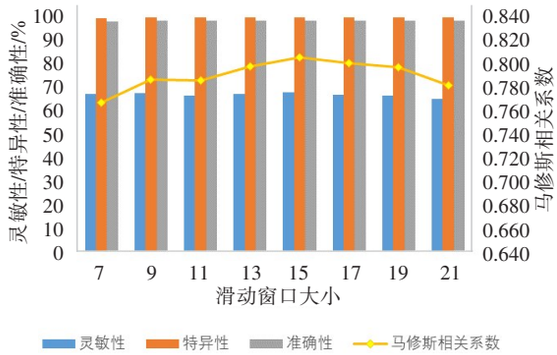

图3 GDP105数据集上基于PSSM特征的滑动窗口结果对比

Figure 3 Comparison of sliding window results based on PSSM features in GDP105 data set

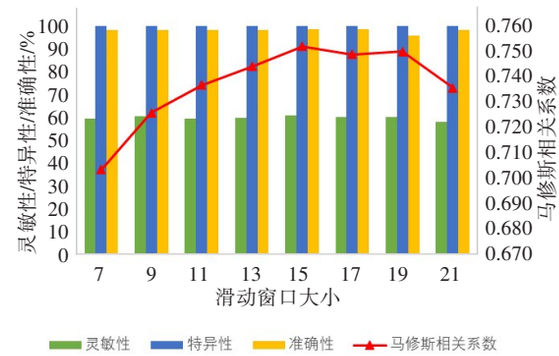

图4 GDP82数据集上基于PSSM特征的滑动窗口结果对比

Figure 4 Comparison of sliding window results based on PSSM features in GDP82 data set

2.3 性能分析

滑动窗口取不同值时,模型在GDP105数据集和GDP82数据集上经过5重交叉验证之后的性能表现见图3和图4。可以看出,Sen、Acc、MCC 3个评价指标先上升后下降。当滑动窗口大小为15时,综合评价指标MCC的值最高,即分类模型的性能最好。

在GDP105数据集上将本研究的方法与石大宏等[27]的方法、NsitePred[19]、SVMPred[18]的模型性能作对比,结果如表3所示。

表3 在GDP105数据集上与其他方法的性能比较

Table 3 Performance comparison with other methods in GDP105 data set

预测方法	Sen/%	Spe/%	Acc/%	MCC
本文方法	67.72	99.91	98.55	0.806
石大宏等[27]	65.25	99.18	97.78	0.700
NsitePred[19]	64.60	99.10	97.60	0.675
SVMPred[18]	62.30	98.90	97.70	0.655

从性能比较结果来看,本文研究的方法在Sen、Spe、Acc和MCC方面表现较好,分别为67.72%、99.91%、98.55%、0.806。与石大宏等[27]的方法相比,本文方法在Sen、Spe、Acc、MCC指标上分别提高了2.47%、0.73%、0.77%、0.106。MCC指标是衡量二分类模型性能的一个均衡指标,因此本文方法整体性能表现较佳。

2.4 独立测试

为了进一步证明本文所提出方法的有效性,本文还使用BioLip中的独立测试集GDP14测试该模型的预测性能[23]。实验结果如表4所示。

从表4中的实验结果可知,本文方法在独立测试集上表现最好,Sen指标为58.67%,Spe指标为99.78%,Acc为98.01%,MCC为0.656,这4个评估指

标比石大宏等[27]的方法分别高出2.37%、0.33%、1.01%、0.085，比NsitePred方法分别高出2.97%、1.88%、1.91%、0.120，比SVMPred方法分别高出9.17%、2.18%、2.61%、0.190。这表明了CNMW下采样与SVM结合的方法可以有效地预测蛋白质-GDP结合位点。

表4 独立测试集GDP14的实验结果

Table 4 Experimental results in independent test set GDP14

预测方法	Sen/%	Spe/%	Acc/%	MCC
本文方法	58.67	99.78	98.01	0.656
石大宏等[27]	56.30	99.45	97.00	0.571
NsitePred[19]	55.70	97.90	96.10	0.536
SVMPred[18]	49.50	97.60	95.40	0.466

3 结 语

本研究将CNMW下采样算法和SVM算法相结合，根据从蛋白质序列中提取的信息来预测GDP结合位点。首先从数据集中的蛋白质序列提取序列特征，然后结合每个残基的邻域残基的特征，采用基于镜像残基的可变滑动窗口方法提取每个残基的特征向量，并利用一种结合K-means聚类、NearMiss-2方法和加权下采样的CNMW下采样方法解决样本集的不平衡问题，最后使用SVM进行分类预测。与其他基于序列的预测方法相比，本文方法在两个标准数据集上的5重交叉验证和独立测试集GDP14上都获得了较高的MCC值，这表明本文方法可以有效地预测蛋白质序列中的GDP结合位点，为分子间相互作用研究提供了新的思路。

【参考文献】

[1] Durrant JD, Mccammon JA. Molecular dynamics simulations and drug discovery[J]. BMC Biology, 2011, 9: 71.

[2] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.

[3] Ballester PJ, Mitchell JB. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking [J]. Bioinformatics, 2010, 26(9): 1169-1175.

[4] Seco J, Luque FJ, Barril X. Binding site detection and druggability index from first principles[J]. J Med Chem, 2009, 52(8): 2363-2371.

[5] Heo L, Shin WH, Lee MS, et al. GalaxySite: ligand-binding-site prediction by using molecular docking[J]. Nucleic Acids Res, 2014, 42: W210-W214.

[6] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool [J]. J Mol Biol, 1990, 215(3): 403-410.

[7] Wälti MA, Riek R, Orts J. Fast NMR-based determination of the 3D structure of the binding site of protein-ligand complexes with weak affinity binders[J]. Angew Chem Int Ed Engl, 2017, 56(19): 5208-5211.

[8] Hogeweg A, Sowislok A, Schrader T, et al. An NMR method to pinpoint supramolecular ligand binding to basic residues on proteins [J]. Angew Chem Int Ed Engl, 2017, 56(46): 14758-14762.

[9] Harris R, Olson AJ, Goodsell DS. Automated prediction of ligand-binding sites in proteins[J]. Proteins, 2010, 70(4): 1506-1517.

[10] Nisius B, Sha F, Gohlke H. Structure-based computational analysis of protein binding sites for function and druggability prediction[J]. J Biotechnol, 2012, 159(3): 123-134.

[11] Si JN, Cui J, Cheng J, et al. Computational prediction of RNA-binding proteins and binding sites[J]. Int J Mol Sci, 2015, 16(11): 26303-26317.

[12] Taherzadeh G, Yang Y, Zhang T, et al. Sequence-based prediction of protein-peptide binding sites using support vector machine[J]. J Comput Chem, 2016, 37(13): 1223-1229.

[13] Srivastava A, Kumar M. Prediction of zinc binding sites in proteins using sequence derived information[J]. J Biomol Struct Dyn, 2017, 36(16): 4413-4423.

[14] Yang XX, Wang J, Sun J, et al. SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues[J]. PLoS One, 2015, 10(7): e0133260.

[15] Walia RR, Xue LC, Katherine W, et al. RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins[J]. PLoS One, 2014, 9(5): e97725.

[16] Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures[J]. Nucleic Acids Res, 2009, 37: W413-W416.

[17] Chauhan JS, Mishra NK, Raghava GP. Identification of ATP binding residues of a protein from its primary sequence [J]. BMC Bioinformatics, 2009, 10: 434.

[18] Chen K, Mizianty MJ, Kurgan L. ATPsite: sequence-based prediction of ATP-binding residues[J]. Proteome Sci, 2011, 9(Suppl 1): S4.

[19] Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors[J]. Bioinformatics, 2012, 28(3): 331-341.

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

[21] Yu DJ, Hu J, Huang Y, et al. TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble[J]. J Computat Chem, 2013, 34(11): 974-985.

[22] Fang C, Noguchi T, Yamana H. Simplified sequence-based method for ATP-binding prediction using contextual local evolutionary conservation[J]. Algorithms Mol Biol, 2014, 9(1): 7.

[23] Yang JY, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions[J]. Nucleic Acids Res, 2013, 41(1): 1096-1103.

[24] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein databases search programs[J]. Nucleic Acids Res, 1997, 25(17): 3389-3402.

[25] Mani I. KNN approach to unbalanced data distributions: a case study involving information extraction[C]//ICML Workshop on Learning from Imbalanced Datasets, 2003.

[26] Cortes C, Cortes C, Vapnik V, et al. Support-vector networks[J]. Machine Learning, 1995, 20: 273-297.

[27] 石大宏, 何雪. 序列蛋白质-GDP绑定位点预测[J]. 计算机工程与应用, 2016, 52(13): 55-59.
Shi DH, He X. Sequence protein-GDP binding site prediction[J]. Computer Engineering and Applications, 2016, 52(13): 55-59.

（编辑：薛泽玲）