Hans Journal of Computational Biology 计算生物学, 2022, 12(3), 23-31
Published Online September 2022 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2022.123004

Hans汉斯

基于机器学习算法识别疾病相关的蛋白与金属离子配体的结合残基

邹向辉，冯永娥*

内蒙古农业大学理学院，内蒙古 呼和浩特

收稿日期：2022年7月18日；录用日期：2022年8月18日；发布日期：2022年8月30日

摘　要

在研究疾病发生机制中，蛋白质与配体相互作用扮演着重要的角色。因为许多蛋白质功能的实现需要结合特定的配体，而金属离子配体对蛋白质功能的实现起到重要作用。确定蛋白质中哪些残基与金属离子配体相互作用，可以帮助研究者理解蛋白质-金属离子相互作用的分子机制，也对人类健康和精准医学有重要意义。本文基于机器学习算法，研究疾病相关的蛋白质与三种金属离子配体的结合。我们分别提取3种序列特征：位置特异性打分矩阵、氨基酸组分信息、二肽组分，并使用随机森林算法和支持向量机算法建立了三种金属离子配体结合残基的分类模型。对于Zn^{2+}结合残基在特征融合中最高准确率(Acc)达到了87%，Mg^{2+}结合残基识别的最高准确率(Acc)达到70%，Ca^{2+}结合残基识别的最高准确率(Acc)达到70%。可见我们的模型对三种金属离子配体的结合残基有一定的识别能力。

关键词

金属离子配体，5折交叉检验，位置特异性打分矩阵，随机森林算法

Identifying the Binding Residues between Disease-Associated Proteins and Metal-Ion Ligands Based on Machine Learning Algorithm

Xianghui Zou, Yong'e Feng*

College of Science, Inner Mongolia Agriculture University, Hohhot Inner Mongolia

Received: Jul. 18th, 2022; accepted: Aug. 18th, 2022; published: Aug. 30th, 2022

*通讯作者。

文章引用: 邹向辉, 冯永娥. 基于机器学习算法识别疾病相关的蛋白与金属离子配体的结合残基[J]. 计算生物学, 2022, 12(3): 23-31. DOI: 10.12677/hjcb.2022.123004

Abstract

Protein-ligand interactions play an important role in the pathogenesis of diseases. Many proteins perform their functions by binding to specific ligands, and the binding of protein-metal-ion ligands plays an important role in the realization of protein functions. Identifying which residues in the protein interact with metal-ion ligands can help researchers understand the molecular mechanism of protein-metal ion interaction, and it is important for human health and precision medicine. In this paper, we study the binding of disease-associated proteins to three metal ion ligands based on the machine learning algorithm. We extract three sequence features: position-specific scoring Matrix (PSSM), amino acid component information, dipeptide component. Then, the random forest algorithm and the support vector machine algorithm were used to establish the classification model of the three metal ion ligand-binding residues. Finally, the highest accuracy (Acc) was 87% for the Zn^{2+} binding residues in the feature fusion, the highest Accuracy (Acc) of Mg^{2+} binding residues was 70%, and that of Ca^{2+} binding residues was 70%. These results show that our model has the ability to identify the binding residues of three metal ion ligands.

Keywords

Metal-Ion Ligand, 5-Fold Cross Validation, Position-Specific Scoring Matrix (PSSM), Random Forest (RF)

1. 引言

蛋白质与配体的相互作用本质上就是通过蛋白质序列上的部分残基的相互作用来实现的，这些残基被称为蛋白质的相互作用位点。而蛋白质功能的实现离不开与配体的相互作用，而这其中蛋白质与金属离子配体相结合就发挥着很重要的生物学功能。研究蛋白质与金属离子配体的结合残基，对理解生命活动的机制，探究蛋白质相互作用原理，发现新的药物靶标等相关研究都具有重要的影响和意义。

目前，研究者们对金属离子配体结合残基进行了大量的预测工作[1]。2004 年，Sodhi 等人[2]利用序列特征，结合人工神经网络(Artificial neural network, ANN)方法，在 5 交叉检验下 Zn^{2+}、Fe^{3+}、Ca^{2+}等配体总精度(Accuracy, Acc)达到了 94.5%。2006 年 Lin 等人[3]基于序列的理化特征对 10 种配体结合残基进行识别，支持向量机(Support Vector Machine, SVM)算法下，预测总精度达到了 74.9%以上。2016 年 Jiang 等人[4]基于蛋白质的序列信息，利用信息学和统计学算法对 Ca^{2+}配体的结合残基进行预测，5 交叉检验下得到的预测总精度为 75.0%。2017 年 Cao 等人[5]对 BioLip 数据库中的 Fe^{2+}、Fe^{3+}、Co^{2+}等 10 种金属离子配体的结合残基进行预测识别，应用序列组合特征结合 SVM 算法，5 交叉检验下 MCC 值均达到 0.5 以上，总精度均高于 74.8%。2020 年，Liu 等人[6]应用二面角作为参数结合随机森林(Random Forest, RF)算法，对 10 种离子结合残基进行预测，总精度高于 77%，MCC 值高于 0.55。2021 年，Wang 等人[7]应用能量特征参数结合支持向量机算法，对 10 种离子结合残基进行预测，独立检验最佳精度值高于 92.5%。综上所述，有关蛋白质与各类金属离子配体的结合已经进行了大量的研究，并提供了一些可靠的模型。但目前尚未有一种令人满意的方法来特定的识别疾病相关蛋白质与金属离子配体的结合位点。本文基于此，首先构

建了人类三大疾病相关的蛋白质数据库，并利用其结构信息和 biolip 数据库[8]，建立了与三种金属离子结合残基的数据集，基于该数据集，我们提取了序列保守性 PSSM 特征、氨基酸组分特征，二肽组分特征，并使用随机森林和支持向量机算法对三类金属离子配体结合残基进行预测，取得了较好的结果。

2. 材料与方法

2.1. 数据集

2.1.1. 数据集的构建

本文基于 Uniprot 数据库[9]中注释信息，获得了三类疾病(心血管疾病、神经退行性疾病、癌症)相关的蛋白，然后利用 Biolip 数据库[8]，获得这三类疾病相关蛋白与三种金属离子 Ca^{2+}、Mg^{2+}、Zn^{2+}配体的结合残基信息，并在序列中标记结合残基。为构建非冗余数据集，我们对序列进行了筛选，首先剔除序列长度不足 50 个氨基酸、其三维结构分辨率大于 3 Å，以及序列一致性高于 30%的蛋白质链。最后获得三种金属离子配体与三类疾病相关蛋白的结合残基数据，列在表 1 中。

Table 1. Datasets of three metal ion ligands
表 1. 三种金属离子配体数据集

Ligands	Chains	P	N
Zn^{2+}	305	1699	83,828
Mg^{2+}	245	1079	88,774
Ca^{2+}	184	1305	62,590

Notes：Chains 表示与金属离子配体结合的蛋白质链数；P 表示金属离子配体的结合残基数量；N 表示金属离子配体的非结合残基数量。

2.1.2. 不平衡数据集的处理

鉴于数据集中正集(P)数量远远小于负集(N)，严重的数据倾斜普遍存在于蛋白质——配体结合残基预测中，为避免不平衡数据集对模型预测性能的影响，我们采取随机采样的手段，在非结合位点数据集中随机选取了与正集数量相等的序列片段作为负集；同时为了确保预测结果的科学性，我们进行了采样的最终结果取 N/P(取整)次预测结果的平均值。

2.2. 特征参数的选取

2.2.1. 位置特异性打分矩阵

位置特异性打分矩阵(Position-Specific Score Matrix, PSSM)可以反映蛋白质序列上每个氨基酸的进化保守信息。在生物学上面认为，相互作用位点通常是一些保守的氨基酸，因此在本文中我们选择了 PSSM 作为特征参数。我们首先利用 BLAST 软件包中的 PSI-BLAST 来搜索 Uniref90 数据库来生成 PSSM 文件，迭代次数设为 3，期望值设为 0.001，其它参数均采用默认值。对于每一条蛋白质序列 P 来说，其 PSSM 可表示为如下形式：

$$P = A_1 A_2 A_3 \cdots A_L \tag{1}$$

$$P_{\mathrm{PSSM}} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,20} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ A_{L,1} & A_{L,2} & \cdots & A_{L,20} \end{bmatrix} \tag{2}$$

其中 A_1 表示蛋白质序列的第一个氨基酸残基，A_2 表示第二个氨基酸残基，以此类推，A_L 表示第 L 个氨基酸。PSSM 为一个 $L \times 20$ 的矩阵，其中 20 表示标准氨基酸，L 为该蛋白质序列的长度，矩阵中元素 $A_{i,j}$ 表示序列上第 j 个氨基酸突变为第 i 个氨基酸的得分，得分值越低说明概率越小，得分值越高表示概率越高。

2.2.2. 氨基酸组分

本文选取的 20 种氨基酸组分(Amino acids composition, AAC)信息作为特征参量，对于数据集中的序列我们可以利用 20 种氨基酸组成来表示。

$$A = \left[A_1, A_2, A_3, \cdots, A_{20}\right] \tag{3}$$

$$A_i = \frac{a_i}{\sum a_i} \quad \left(i = 1, 2, 3, \cdots, 20\right) \tag{4}$$

2.2.3. 二肽组分

本文选取的 400 种二肽组分(Dipeptide composition, DC)信息作为特征参量，对于数据集中的序列我们可以利用 20 种氨基酸组成来表示。

$$DC = \left[DC_1, DC_2, DC_3, \cdots, DC_{400}\right] \tag{5}$$

$$DC_i = \frac{DC_i}{\sum DC_i} \quad \left(i = 1, 2, 3, \cdots, 400\right) \tag{6}$$

2.3. 算法

2.3.1. 方差分析

方差分析(Analysis of Variance，简称 ANOVA)又称为“F 检验”或“变异数分析”[10]，这种方法常用于两个及两个以上样本均数差别的显著性检验。我们选用单因素方差分析，公式(7)计算多组样本均数的显著性差异：

$$MS_T = MS_B + MS_W \tag{7}$$

其中 MS_T 代表总均方，MS_B 代表组间均方，MS_W 代表组内均方。统计值 F 值是组间均方和组内均方的比值，为了消除因各组样本数不同而产生的影响，F 值的计算如公式(8)表示

$$F = \frac{MS_B}{MS_W} \tag{8}$$

F 值大就说明处理之间差异比较明显，误差项小就说明试验的精度较高。一般 F 值越大，P 值越小。统计学上规定，一般当 P 值小于 0.05 时，可以说各组样本间存在差异，当 P 值小于 0.01 时，则说明各组样本之间存在着显著的差异。本文利用方差分析对三种金属离子配体的结合残基和非结合残基进行了氨基酸分布是否具有差异显著的分析。

2.3.2. 随机森林算法(RF)

随机森林(Random Forest，简称 RF)算法是 Leo Breiman 在 2001 年提出的一种分类预测模型[11]，是由许多单棵分类回归树组合而成的，一棵分类回归树就是一个分类器，最后的决策结果由投票法决定。它的基本思想是将很多弱分类器集成一个强分类器。随机森林算法是一种通过自助法采样来构造多个分类器的组合分类器[12]。它通过在各个节点处随机选择特征进行分支，这样可以最小化各棵分类树之间的相关性，从而提高分类的精度，所以随机森林算法已经被广泛地应用到分类以及模式识别等问题中[13] [14]。

随机森林有两个重要的参数，一个是单棵决策树每个节点处分裂时所选用的候选特征参数的个数 m，另一个是随机森林中决策树的棵数 k (k = 500)。用随机森林分类器对新的数据进行判别与分类，按照树分类器进行投票，最后由投票法决定分类结果。随机森林通过在每个节点处随机选择特征进行分支，这样可以最小化各棵分类树之间的相关性，提高分类的精确性。随机森林算法不会出现过度拟合现象、分类效率也很高，而且能够快速处理大样本数据，同时需要调整的参数也比较少，能更好的估计哪个特征在分类中更重要。

2.3.3. 支持向量机(SVM)

支持向量机(Support Vector Machine，简称 SVM)是由 Vapink 等人在 1995 年所提出的一种基于统计学习的机器学习算法，它的基本模型是定义在特征空间上的间隔最大化线性分类器。SVM 在各领域内的二分类和多分类问题中都有应用[15]。在 SVM 中，数据会通过核函数将低维线性不可分的数据映射到高维空间中，使得原本不可分的数据变得线性可分。最后通过最优化算法求得数据集的几何间隔最大的分离超平面。本文选取了径向基核函数(Radial Basis Function，简称 RBF)来训练模型。通过台湾大学 Lin Chih-Jen 开发的 LIBSVM3.21 软件包[16]，搜寻最优的参数，来实现特征参数的最优化和预测。

2.4. 评价指标

目前，预测算法性能检验常用的方法主要有独立检验(independent test)和 K-折交叉检验(K-fold cross-validation test)。本文采用 5 折交叉检验，即将数据集随机分为 5 个子集合，依次从中取出一个子集作为测试集，而将剩余的 4 个子集合则作为训练集，此过程一共循环 5 次。

对于任何预测算法性能的评价，主要是保证该预测算法能对属于同一数据域的新样本具有推广性能。本文采用了四种评估指标来评估模型的性能，精确度(Precision, Pre)、召回率(Recall, Rec)、预测总精度(Accuracy, Acc)、马修斯相关系数(Mathews correlation coefficient, Mcc)，定义如下。

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{9}$$

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{10}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \tag{11}$$

$$\mathrm{Mcc}=\frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})}} \tag{12}$$

3. 结果与讨论

3.1. 结合残基与非结合残基处氨基酸分布的差异性分析

本文统计了金属离子 Zn^{2+}，Mg^{2+}，Ca^{2+}结合残基与非结合残基处的 20 种氨基酸的分布，并利用方差分析，公式(7)~(8)对三种金属离子配体的结合残基与非结合残基处的氨基酸的分布进行了差异是否显著的分析，结果如表 2 所示，Ca^{2+}结合残基与非结合残基处有显著差异的氨基酸有 16 个，分别是氨基酸 A、C、D、E、G、I、H、K、L、N、Q、P、R、S、V；在 ZN^{2+}结合残基与非结合残基处有显著差异的氨基酸有 13 个，分别是氨基酸 A、C、E、H、I、K、L、N、Q、R、S、T、V；在 Mg^{2+}结合残基与非结合残基处有显著差异的氨基酸有 9 个，分别是氨基酸 D、E、G、H、K、L、P、Q、R。可见三种金属离子的结合残基与非结合残基处氨基酸的分布是有显著差异的，基于此接下来利用序列特征识别这些结合残基。

Table 2. Amino acids with significant differences in the binding and non-binding residues for 3 metal ion ligands
表 2. 三种金属离子配体的结合残基与非结合残基处有显著差异的氨基酸

Zn^{2+}	Mg^{2+}	Ca^{2+}
Residues of P-Value less than 0.01 ($P < 0.01$)		
A	D	A
C	E	C
E	G	D
H	H	E
I	K	G
K	L	H
L	P	I
N	Q	K
Q	R	L
R		N
S		P
T		Q
V		R
		S
		V

3.2. 随机森林(RF)算法的预测结果

考虑到蛋白质与金属离子配体相互作用时，并不只与某一特定结合残基发生相互作用，也受到其周围残基的影响，所以我们选用滑动窗口的方法截取一定长度的序列片段，若片段中心为金属离子配体的结合残基，则定义该片段为正集片段，反之则为负集片段。为保证蛋白质上的每一个残基都能出现在窗口中心，我们分别在蛋白质链的两端加上$(L-1)/2$ 个伪氨基酸 X，L 为所选取的窗口长度。

下面是利用随机森林算法在单特征(PSSM, AAC, DC)以及特征融合后 5 折交叉检验的预测结果列在表 3 中。其中应用 20 种氨基酸组分(AAC)在 Zn^{2+}、Mg^{2+}、Ca^{2+}结合残基与非结合残基识别中最佳识别窗口分别为 11、9、7；400 种二肽组分(DC)在 Zn^{2+}、Mg^{2+}、Ca^{2+}结合残基与非结合残基识别中最佳识别窗口分别为 7、7、5；特异性打分矩阵(PSSM)在 Zn^{2+}、Mg^{2+}、Ca^{2+}结合残基与非结合残基识别中最佳识别窗口分别为 5、7、9；在组合特征最佳窗口取 9 的情况，精度有一定程度的提高，可见以上特征融合还是有助于识别这些结合残基的。

3.3. 支持向量机(SVM)算法的预测结果

在数据集中，我们再次选取 AAC (最佳窗口均为 11)和 DC (最佳窗口均为 7)作为特征参数，用支持向量机算法进行预测，在 5 折交叉检验下，预测结果列在表 4。

鉴于特征融合计算量大，故在 SVM 预测算法中没有进行特征的融合，但对比表 3 的结果，发现在相同的特征参数(AAC、DC)下，随机森林算法(RF)比支持向量机算法更适合识别这三种金属离子配体的结合残基。

Table 3. Prediction results based on RF in 5-fold cross-validation
表 3. 5-折交叉下随机森林算法的预测结果

Ligands	*Features*	*Pre*	*Rec*	*Acc*	*MCC*
Zn^{2+}	PSSM	0.92	0.77	0.85	0.714
	AAC	0.81	0.77	0.79	0.590
	DC	0.76	0.86	0.80	0.598
	AAC + PSSM	0.92	0.80	0.87	0.739
	DC + PSSM	0.91	0.77	0.85	0.707
Mg^{2}	PSSM	0.72	0.63	0.69	0.381
	AAC	0.58	0.51	0.58	0.164
	DC	0.60	0.55	0.59	0.184
	AAC + PSSM	0.72	0.63	0.69	0.385
	DC + PSSM	0.72	0.62	0.69	0.378
Ca^{2+}	PSSM	0.73	0.63	0.70	0.397
	AAC	0.67	0.59	0.65	0.300
	DC	0.65	0.63	0.64	0.287
	AAC + PSSM	0.74	0.64	0.70	0.413
	DC + PSSM	0.73	0.63	0.70	0.394

Table 4. Prediction results based on SVM in 5-fold cross-validation
表 4. 5-折交叉下支持向量机算法的预测结果

Ligands	*Features*	*Pre*	*Rec*	*Acc*	*MCC*
Zn^{2+}	AAC	0.81	0.76	0.79	0.587
	DC	0.85	0.63	0.76	0.534
Mg^{2}	AAC	0.60	0.53	0.59	0.182
	DC	0.76	0.12	0.54	0.148
Ca^{2+}	AAC	0.68	0.53	0.64	0.280
	DC	0.79	0.21	0.58	0.226

3.4. 三种金属离子配体结合位点的预测结果

金属离子配体的结合残基的准确识别有助于理解蛋白质的功能，所以我们进一步应用随机森林和支持向量机两种算法对三种金属离子的结合残基进行 3 分类的识别，滑动窗口均取 11 时分类效果最好，结果列在表 5 中。

Table 5. Prediction of binding sites of 3 metal ion ligands
表 5. 三种金属离子配体结合位点的预测结果

Method	*Ligands*	*Features*	*Pre*	*Rec*	*Acc*	*MCC*
RF	Zn^{2+}	AAC	0.77	0.84	0.73	0.588
	Mg^{2}		0.68	0.53		
	Ca^{2+}		0.71	0.76		

Continued

RF	Zn^{2+}	DC	0.81	0.89	0.82	0.719
	Mg^{2}		0.79	0.69		
	Ca^{2+}		0.84	0.82		
SVM	Zn^{2+}	AAC	0.80	0.82	0.72	0.572
	Mg^{2}		0.66	0.56		
	Ca^{2+}		0.66	0.73		
SVM	Zn^{2+}	DC	0.83	0.88	0.81	0.710
	Mg^{2}		0.74	0.71		
	Ca^{2+}		0.83	0.81		

以上结果可见，二肽组分(DC)特征更有利于识别三种金属离子结合残基。对比两种算法的结果，可以看出对于三种金属离子配体的结合残基，随机森林(RF)算法要比支持向量机(SVM)算法识别效果好。

4. 结论

本文研究的蛋白质是源于 Uniprot 库中与人类疾病相关的三类蛋白，中间剔除了结合后有突变残基的蛋白、还有长度不足 50 个残基的蛋白以及没有对应三维结构信息的蛋白，最后获得的蛋白质数量有限，所以本文只研究了其结合位点较多的三种金属离子配体的结合。首先，通过方差分析发现三种金属离子的结合残基与非结合残基处氨基酸的分布存在显著的差异性，然后，基于氨基酸序列分别提取 3 种特征(PSSM, AAC, DC)应用两种算法实施分类。在三种金属离子结合残基与非结合残基识别中，应用随机森林算法(RF)，氨基酸组分(AAC)结合特异性打分矩阵(PSSM)特征预测正确率较高，Zn^{2+}结合残基的预测总精度(Acc)最高达到 87%。而在三种金属离子配体结合残基的分类识别中，利用随机森林算法(RF)和二肽组分(DC)特征结合，取得总精度(Acc)最高为 82%。可见，该模型对于疾病相关蛋白与金属离子配体的结合残基还是有一定的识别能力的。下一步随着公共数据库的扩增，我们获得更多疾病相关的蛋白，将开展多种金属离子配体的研究。

基金项目

感谢匿名的评审专家对本文给出的宝贵意见，同时感谢国家自然科学基金专项项目(62141204)，内蒙古自治区研究生教改项目(YJG20191012908)对本论文的资助。

省略语表

SVM (Support Vector Machine)：支持向量机；
PSSM (Position-Specific Score Matrix)：位置特异性打分矩阵；
RF (Random Forest)：随机森林；
ANN (Artificial neural network)：人工神经网络；
ANOVA (Analysis of Variance)：方差分析；
AAC (Amino acids composition)：氨基酸组分；
DC (Dipeptide composition)：二肽组分；
Acc (Accuracy)：准确率；
MCC (Matthew’s correlation coefficient)：马修斯相关系数；

Pre(Precision)：精确度；

Rec(Recall)：召回率。

参考文献

[1] 张晓瑾. 基于 GBM 算法识别蛋白质中金属离子配体的结合残基[D]: [硕士学位论文]. 呼和浩特: 内蒙古工业大学, 2019.

[2] Sodhi, J.S., Bryson, K., McGuffin, L.J., *et al*. (2004) Predicting Metal-Binding Site Residues in Low-Resolution Structural Models. *Journal of Molecular Biology*, **342**, 307-320. https://doi.org/10.1016/j.jmb.2004.07.019

[3] Lin, H.H., Han, L.Y., Zhang, H.L., *et al*. (2006) Prediction of the Functional Class of Metal-Binding Proteins from Sequence Derived Physicochemical Properties by Support Vector Machine Approach. *BMC Bioinformatics*, **7**, S13. https://doi.org/10.1186/1471-2105-7-S5-S13

[4] Jiang, Z., Hu, X.Z. Geriletu, G., *et al*. (2016) Identification of Ca^{2+}-Binding Residues of a Protein from Its Primary Sequence. *Genetics and Molecular Research*, **15**, gmr.15027618. https://doi.org/10.4238/gmr.15027618

[5] Cao, X.Y., Hu, X.Z., Zhang, X.J., *et al*. (2017) Identification of Metal Ion Binding Sites Based on Amino Acid Sequences. *PLOS ONE*, **12**, e0183756. https://doi.org/10.1371/journal.pone.0183756

[6] Liu, L., Hu, X.Z., Feng, Z.X., *et al*. (2020) Recognizing Ion Ligand-Binding Residues by Random Forest Algorithm Based on Optimized Dihedral Angle. *Frontiers in Bioengineering and Biotechnology*, **8**, Article 493. https://doi.org/10.3389/fbioe.2020.00493

[7] Wang, S., Hu, X.Z., Feng, Z.X., *et al*. (2021) Recognition of Ion Ligand Binding Sites Based on Amino Acid Features with the Fusion of Energy, Physicochemical and Structural Features. *Current Pharmaceutical Design*, **27**, 1093-1102. https://doi.org/10.2174/1381612826666201029100636

[8] Yang, J.Y., Roy, A. and Yang, Z.Y. (2013) BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand-Protein Interactions. *Nucleic Acids Research*, **41**, D1096-D1103. https://doi.org/10.1093/nar/gks966

[9] Bateman, A., Martin, M.-J., Orchard, S., *et al.* (2020) UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480-D489.

[10] Kou, G.S. and Feng, Y.E. (2015) Identify Five Kinds of Simple Super Secondary Structures with Quadratic Discriminant Algorithm Based on the Chemical Shifts. *Journal of Theoretical Biology*, **380**, 392-398. https://doi.org/10.1016/j.jtbi.2015.06.006

[11] Breiman, L. (2001) Random Forests, Machine Learning 45. *Journal of Clinical Microbiology*, **2**, 199-228. https://doi.org/10.1023/A:1010933404324

[12] Li, Z.C., Lai, Y.H., Chen, L.L., *et al*. (2012) Identification of Human Protein Complexes from Local Sub-Graphs of Protein-Protein Interaction Network Based on Random Forest with Topological Structure Features. *Analytica Chimica Acta*, **718**, 32-41. https://doi.org/10.1016/j.aca.2011.12.069

[13] Walsh, E.S., Kreakie, B.J., Cantwell, M.G. and Nacci, D. (2017) A Random Forest Approach to Predict the Spatial Distribution of Sediment Pollution in an Estuarine System. *PLOS ONE*, **12**, e0179473. https://doi.org/10.1371/journal.pone.0179473

[14] Yang, L., Wu, H., Jin, X., *et al*. (2020) Study of Cardiovascular Disease Prediction Model Based on Random Forest in Eastern China. *Scientific Reports*, **10**, Article No. 5245. https://doi.org/10.1038/s41598-020-62133-5

[15] Sun, C.Z. and Feng, Y.E. (2021) Identify Disordered Regions of Intrinsically Disordered Proteins by Multi-Features Fusion. *Current Bioinformatics*, **16**, 1126-1132. https://doi.org/10.2174/1574893616666210308102552

[16] Chang, C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), **2**, Article 27. https://doi.org/10.1145/1961189.1961199