Hans Journal of Computational Biology 计算生物学, 2018, 8(1), 24-32
Published Online March 2018 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2018.81004

Hans汉斯

Identification of Protein Phosphorylation Sites by Diversity Increment Feature Selection Technique

Shisai Hu, Zhen Liang, Yuxiang Chen, Ying Zhang*, Jun Lv*

College of Science, Inner Mongolia University of Technology, Hohhot Inner Mongolia
Email: *yzhang@imut.edu.cn, *lujun@imut.edu.cn

Received: Apr. 25th, 2018; accepted: May 15th, 2018; published: May 22nd, 2018

Abstract

Phosphorylation is one of the most important protein post-translational modifications and plays important roles in numerous biological processes by significantly affecting proteins' structure and dynamics. The development of computational biological methods for the accurate identification of phosphorylation sites helps to our understanding of key signal transduction mechanisms. In this paper, a kinase independent phosphorylation site identification model was presented, called FSID_PhSite. The model is featured by component of k-spaced amino acid pairs and the position conservation of residues surrounding the phosphorylation sites. Applying diversity incremental feature selection technique to feature selection and inputting the selected features into the support vector machine algorithm for recognition, when the ratio of positive and negative samples is 1:1, on independent testing dataset validation, the accuracy of identification for serine, threonine and tyrosine sites is 84.34%, 82.32% and 68.89%, respectively. The results were superior to the existing kinase independent phosphorylation sites identification model.

Keywords

Protein Phosphorylation Site, Feature Selection Based on Increment of Diversity, Support Vector Machine

用多样性增量特征选择技术识别蛋白质磷酸化位点

胡世赛，梁　珍，陈宇翔，张　颖*，吕　军*

内蒙古工业大学理学院，内蒙古　呼和浩特
Email: *yzhang@imut.edu.cn, *lujun@imut.edu.cn

*通讯作者。

文章引用: 胡世赛, 梁珍, 陈宇翔, 张颖, 吕军. 用多样性增量特征选择技术识别蛋白质磷酸化位点[J]. 计算生物学, 2018, 8(1): 24-32. DOI: 10.12677/hjcb.2018.81004

收稿日期：2018年4月25日；录用日期：2018年5月15日；发布日期：2018年5月22日

摘　要

磷酸化是最重要的蛋白质翻译后修饰之一，在许多细胞过程中扮演重要角色。发展磷酸化位点精确识别的计算生物学方法，有助于对磷酸化信号转导机制的理解。本文给出一种激酶无关的磷酸化位点识别模型，称为FSID_PhSite。模型以k间隔氨基酸对组分和位置保守氨基酸组分为特征，应用多样性增量特征选择技术进行特征筛选，将选出的特征输入到支持向量机算法进行识别。在正负样本数之比为1:1的情形下，对磷酸化丝氨酸、苏氨酸和酪氨酸在独立测试集检验，识别精度分别达到84.34%、82.32%和68.89%。结果优于现有的激酶无关磷酸化位点识别模型。

关键词

蛋白质磷酸化位点，多样性增量特征选择，支持向量机

1. 引言

磷酸化是最常见的真核生物蛋白质翻译后修饰之一，是一个能量依赖的化学过程。在广泛的细胞过程中，磷酸化参与了转录调控、DNA 修复、代谢、免疫反应、环境应激反应和细胞运动等。磷酸化过程是在特异的磷酸激酶的催化下，高能磷酸盐供体 ATP/GTP 等的末端磷酸基团加到特异蛋白链的特异受体氨基酸分子底物上，特异氨基酸的磷酸化改变了这一蛋白的酶活性。一条蛋白质链的磷酸化一般只发生在丝氨酸(serine, S)，苏氨酸(threonine, T)或酪氨酸(tyrosine, Y)这三个残基上。约 30%~50%的真核蛋白质要经历磷酸化过程[1]。磷酸化/去磷酸化在不同细胞组织存在的广泛性，以及它与疾病的密切相关性，使得蛋白质磷酸化问题成为分子生物学研究的重要课题。

近年来，随着实验技术的不断提升，蛋白质翻译后修饰位点数据不断扩充，极大地推动了蛋白质翻译后修饰的研究进展。然而实验方法往往费时费力，成本较高，因此大大推动了高效、精准预测翻译后修饰位点的计算方法的发展。对磷酸化位点预测的计算生物学模型大体可分为三类，一是激酶特异模型[2] [3] [4] [5]，二是物种或组织特异模型[6] [7] [8]，三是物种和激酶无关的模型[9] [10] [11] [12] [13]。特征信息来源一般是，底物序列片段的残基位置保守性，给定窗口残基组分或关联组分，进化保守性等。最近，Audagnotto 和 Dal Peraro [14]对蛋白质翻译后修饰的生物信息学预测工具进行了一个综述，给出了这些模型的 Web 服务器链接，方便研究者查询。由于一个蛋白质中，被修饰的位点是少数，与非修饰位点数相比相差悬殊，因此多数模型的预测精度均是在设定正负集样本数之比为 1:1 情形下给出的。现有模型普遍存在输入特征维数高的缺点，进而致使模型出现严重的过学习情况。尤其是物种和激酶无关的模型，输入特征偏多且预测精度偏低[9] [10] [11] [12] [13]。

随着被研究问题的复杂性的增加，特征向量的维数越来越高，以期获得更高的预测精度。但是，高维特征将导致对样本的过拟合进而导致结果的泛化能力降低。因此，应用特征选择技术进行数据分析和特征优化越来越受到人们的关注。Drotár 等[15]在 8 个二分类的生物医学数据集上，比较了 10 个最先进

的特征选择方法，发现基于熵的特征选择方法(information gain, IG) [16]具有最高的稳定性，而最小冗余最大相关(Minimal Redundancy Maximal Relevance, mRMR)方法[17]具有最高的预测精度。2016 年，Zou 等提出称为最大相关最大距离(Max-Relevance-Max-Distance, MRMD)的特征选择方法[18]。在 2017 年，我们提出了一个新的特征选择方法——多样性增量特征选择(Feature Selection based on Increment of Diversity, FSID) [19]。我们将该方法应用于蛋白质柔性/刚性分类预测[19]问题中，发现 FSID 方法具有高效的特征降维能力，优于 IG 和 MRMD 方法。

蛋白质磷酸化位点识别是一个典型的高维特征问题。因此，采用特征选择技术实现特征降维，是实现精确识别磷酸化位点的可行路线之一。

在本文中，我们提出一个新的与激酶无关的磷酸化位点识别模型 FSID_PhSite。在一个较大的蛋白质磷酸化位点注释数据集上，我们以 k 间隔氨基酸对组分和磷酸化底物片段位置保守残基组分为特征源，采用 FSID 方法进行特征选择，并结合支持向量机算法进行识别，获得了较高的识别精度。

2. 材料与方法

2.1. 数据集

本文数据集分为两部分：训练集和独立检验集。两部分数据均来自 Zhao 等文献[12]的补充材料。其中训练集中的正样本，即实验上确定的磷酸化位点，来源于 Phospho.ELM 数据库 8.1 版本(2008 年 8 月 12 日发布) [20]。训练集包括 5725 个蛋白质，其中磷酸化丝氨酸(serine, S)位点 12373 个、磷酸化苏氨酸(threonine, T)位点 2525 个，磷酸化酪氨酸(tyrosine, Y)位点 1826 个，这些位点组成训练集的正集。训练集中的负样本也来源于这 5725 个蛋白质。一个蛋白质序列中除磷酸化位点外，其余与任一磷酸化位点间距在 50 个氨基酸以上的 S/T/Y 残基，均被认为是非磷酸化位点，该负样本选取方案也被 Biswas 等[11]所采用。由于非磷酸化位点数远大于磷酸化位点数，Zhao 等[12]采取分别选择了十组与磷酸化位点等量的非磷酸化位点作为负集。也即，一个训练正集对十个等量的训练负集。

为了公平地比较不同预测模型的性能，Zhao 等[12]还收集到一个新的独立检验集。该独立检验集取自 Phospho.ELM 数据库 2008 年 8 月 12 日之后的新增数据。经去冗余处理后，将独立检验集中的蛋白质序列相似性降低到 40%以下。最后，在独立检验集中包含 837 个蛋白质，其中磷酸化丝氨酸位点 1450 个、磷酸化苏氨酸位点 835 个、磷酸化酪氨酸位点 286 个。独立检验集中负集的提取方式与训练集中负集的提取方式一致。

2.2. 特征提取方法

2.2.1. k 间隔氨基酸对组分

以 S/T/Y 位点为中心两侧各截取 13 个残基，得到由 27 个残基组成的信息富含片段。其中正集片段第 14 位为磷酸化 S/T/Y 残基，而负集片段第 14 位为非磷酸化 S/T/Y 残基。对于处于蛋白质 N 端或 C 端的 S/T/Y 残基，当两侧不足 13 个残基时，用符号“O”补齐。显然，我们期望这个由 27 个残基组成片段，能够成为识别中心残基是否被磷酸化的充足信息源。

k 间隔氨基酸对组分，反映出待测位点附近一对残基间的关联特征，随着 k 值的由小到大改变，分别表现出近程到远程的关联。k 间隔氨基酸对组分是对一个蛋白质序列较充分的表示，尤其对于短的片段。k 间隔氨基酸对组分的序列信息编码方案，已成功用于多个蛋白质功能位点识别模型中，诸如磷酸化位点预测[12]，柔性位点预测[19]等等。

对于每个 k 的取值(本文取 k = 0, 1, 2, 3, 4, 5；k 是一对氨基酸中间间隔的残基数)，由长度为 27 的残基片段，可提取 441 种(AA, AC, AD, …, OO)氨基酸对组分。显然，k 值越大，所考虑的关联情形就越多，对序列片段的表示也就越充分，但随之而来的特征维数也大幅增高。本文中，我们考虑取 k 值最大到 5

时，特征向量的总维数达到 2646 维。不难推断，这 2646 个特征中的绝大部分应该是类别无关的和冗余的。如果将这些特征均输入分类器用于分类，一来将降低分类器的灵敏度，二来将出现严重的过学习现象。因此，我们需应用特征选择技术对这些特征进行筛选，从中选出少数类别相关的关键特征。

2.2.2. 位置保守的氨基酸组分

k 间隔氨基酸对组分经特征筛选后，必然丢失部分信息，选出少数特征已不再是序列片段的充分表示。为了能够尽可能地弥补丢失信息，我们进一步提取磷酸化底物片段的残基保守性特征作为补充。以 S/T/Y 所在位置为 0 点，上游选取−5:−1 位置，下游选取+1:+5 位置，共 10 个位点。以每个位点上的氨基酸组分为特征，构建 21 × 10 = 210 维特征向量。显然，这些特征中仍存在大量与类别无关的特征，采用同样的特征选择方法对这些特征进行筛选。将选出的位置保守氨基酸组分特征，加入到选出的 k 间隔氨基酸对组分特征中，共同作为识别算法的输入。

2.3. 特征选择方法

特征选择的思想是在尽量少损失精度的前提下，尽可能减少特征数。这有点类似于计算机科学中的文件压缩，压缩比越大，失真率越高，好的压缩算法是最小失真和最大压缩的最佳平衡。近些年，众多新的特征选择技术得以发展[15] [16] [17] [18]。最近，我们研究组发展了一种新的特征选择技术，称为多样性增量特征选择(feature selection based on incremental of diversity, FSID) [19]。FSID 方法是我们在研究蛋白质柔性/刚性分类问题时被提出的[19]。

对于一个两类分类问题 C_i (I = 1, 2)，特征 X 出现在类别 C_1 的频次记为 n，除特征 X 以外的其它特征出现在类别 C_1 的频次记为 $\bar{n}$；类似地，特征 X 出现在类别 C_2 的频次记为 m，其它特征出现在类别 C_2 的频次记为 $\bar{m}$。则特征 X 在 C_1 类别中的多样性量定义为

$$D(X,C_1)=(n+\bar{n})\ln(n+\bar{n})-n\ln(n)-\bar{n}\ln(\bar{n}) \tag{1}$$

同样地，特征 X 在 C_2 类别中的多样性量 D (X, C_2)，以及在混合系统 $C_1 + C_2$ 中的多样性量 D (X, $C_1 + C_2$)可以用同样的方式分别定义为

$$D(X,C_2)=(m+\bar{m})\ln(m+\bar{m})-m\ln(m)-\bar{m}\ln(\bar{m}) \tag{2}$$

和

$$\begin{aligned}D(X,C_1+C_2)=&(n+m+\bar{n}+\bar{m})\ln(n+m+\bar{n}+\bar{m})\\&-(n+m)\ln(n+m)\\&-(\bar{n}+\bar{m})\ln(\bar{n}+\bar{m})\end{aligned} \tag{3}$$

特征 X 在 C_1 和 C_2 类别之间的多样性增量(increment of diversity, ID)定义为

$$ID(X)=D(X,C_1+C_2)-D(X,C_1)-D(X,C_2) \tag{4}$$

容易证明，$ID(X)$满足相似性度量的两个基本条件，非负性和对称性。由多样性增量的定义看出，当某个特征 X 在 C_1 和 C_2 两个类别中出现的频次差异越大时，$ID(X)$的值越大，相反，频次差异越小时，$ID(X)$的值越小。如果特征 X 是类别无关的，那么一般特征 X 在两个类别中出现的频次应几乎无差别，此时 ID(X)的值接近 0。因此，$ID(X)$就可作为特征 X 是否与类别相关的度量，或者，$ID(X)$是针对特征 X 对两个类别 C_1 和 C_2 的相似度的度量。这就是说，如果 $ID(X) > ID(Y)$，表明特征 X 与类别相关性要强于特征 Y。当这种类别相关性的强度达到我们的预期(ID_0)时，即当 $ID(X) > ID_0$，特征 X 被选择，ID_0 为特征选择阈值。这个特征选择方案被我们称为多样性增量特征选择技术(FSID)。

2.4. 分类识别算法

本文分类识别算法采用支持向量机。支持向量机是一种二类分类器，其基本型定义为特征空间上的分类超平面与支持向量间隔最大的线性分类器，采用核函数技术，支持向量机可将原特征空间的非线性分类问题，隐映射到高维空间转换为线性可分问题。本文支持向量机算法的实现采用 R 语言“e1071”包中的 svm [21]函数完成。核函数采用径向核函数，参数 c 和 γ 分别采用默认值。

2.5. 分类识别性能评价指标

分类性能采用如下四个指标度量，分别是敏感性(Sensitivity, Sn)，特异性(Specificity, Sp)，总精度(Accuracy, ACC)和马氏相关系数(Matthews correlation coefficient, MCC)。这些量定义如下

$$\begin{aligned} &\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}, \\ &\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \end{aligned} \tag{5}$$

这里分别为 TP 表示真阳性(真实磷酸化位点被预测为磷酸化位点的数目)，TN 表示真阴性(真实非磷酸化位点被预测为非磷酸化位点的数目)，FP 表示假阳性(真实非磷酸化位点被预测为磷酸化位点的数目)，FN 表示假阴性(真实磷酸化位点被预测为非磷酸化位点的数目)。

上面四个分类性能评价指标分别表明了一个预报器的四个不同方面的性能。Sn 是在全体正样本中能够被正确预测为正样本的频率，它用来衡量一个预报器识别正样本的能力。类似地，Sp 是用来衡量一个预报器识别负样本的能力。ACC 测度正确识别全部样本的能力。MCC 是预测性能的一个最佳平衡测度。MCC 的取值范围是[−1, +1]。MCC = 0 表明预报器实际执行了一个随机猜测，也即它的预测结果与样本的真实分类标签不相关。MCC = ±1 表明预报器是完美的。同时给出一个预报器的四个性能指标，可以较全面地反映出预报器的输出性能。

3. 结果与讨论

3.1. FSID 特征选择结果

所有的特征选择过程均在训练集上完成。一个训练正集分别对十个等量的训练负集，组成 10 组训练集。如果某个特征在 10 组训练集上所得 ID 值均大于阈值 ID_0，则该特征被选出，被选出的特征输入到 SVM 中进行磷酸化/非磷酸化位点的分类识别。阈值 ID_0 由识别总精度最大化定出。

经上述操作后，对于 k 间隔氨基酸对组分特征，分别选出用于识别丝氨酸的特征 72 个，识别苏氨酸的特征 14 个，识别酪氨酸的特征 14 个。对于位置保守氨基酸组分特征，分别选出用于识别丝氨酸的特征 34 个，识别苏氨酸的特征 27 个，识别酪氨酸的特征 26 个。这样，最后得到磷酸化丝氨酸的识别特征共 72 + 34 = 106 个，苏氨酸的识别特征共 14 + 27 = 41 个，酪氨酸的识别特征共 14 + 26 = 40 个。表 1 给出了三类磷酸化残基对应选出的位置保守氨基酸组分特征。

由表 1 可见，磷酸化丝氨酸位点附近的位置保守片段为 RXR[RS]DS[PSD][ES][ES]SS，磷酸化苏氨酸位点附近的位置保守片段为 R[CS]R[PS]XTP[PT]TTP，磷酸化酪氨酸位点附近的位置保守片段为 XD[DE][DP]DY[EY]Y[PVY]YY。

这三个位置保守片段有一个显著的共同特点：磷酸化丝氨酸位点的紧邻下游连续保守地出现多个丝氨酸，磷酸化苏氨酸位点的紧邻下游连续保守地出现多个苏氨酸，磷酸化酪氨酸位点的紧邻下游也连续保守地出现多个酪氨酸。据此我们推测，这可能是磷酸化过程的一个有效的容错机制。也即，每个磷酸

Table 1. Features of the position conservation selected by the FSID method
表 1. 由 FSID 方法选择的位点保守性特征

位点		−5	−4	−3	−2	−1	1	2	3	4	5
保守氨基酸*	丝氨酸	**R**		**R**	**R**	**D**	**P**	**E**	**E**	**S**	**S**
					S		**S**	**S**	**O**	**O**	**O**
					L		**D**	L	**S**		
							H	F	V		
							K		L		
							L				
							N				
							A				
							C				
							T				
	苏氨酸	**R**	**C**	**R**	**P**		**P**	**P**	**T**	**T**	**P**
			S		**S**		C	**T**	**O**	**O**	**O**
							G	**O**		W	T
							K	L		L	
							L	F			
							A				
							T				
	酪氨酸		**D**	**D**	**D**	**D**	**E**	**Y**	**P**	**Y**	**Y**
			L	**E**	**P**		**Y**	**O**	**V**	**O**	**O**
				I			K		**Y**	I	
				L			P		**O**	L	
							L				

*注：加粗显示的为磷酸化位点保守性氨基酸残基，其他为非磷酸化位点保守性氨基酸残基。

化位点的紧邻下游均会出现同一种氨基酸，作为替补磷酸化位点，在特异的磷酸激酶的催化下，高能磷酸盐供体 ATP/GTP 的末端磷酸基团一旦没有及时识别需要磷酸化的特异氨基酸，那么其后的相同氨基酸则会代替特异氨基酸进行磷酸化过程。

3.2. FSID_PhSite 模型的识别结果

在独立测试集上，FSID_PhSite 模型对丝氨酸、苏氨酸和酪氨酸的 10 组检验结果的平均值分别列于表 2、表 3 和表 4 中。为了比较不同模型的识别性能，我们同时列出了其他 4 个最好的激酶无关模型在同样测试集上的识别结果。这四个模型的结果均取值文献[12]。

由表 2 可见，我们的模型 FSID_PhSite 对磷酸化丝氨酸位点的识别性能分别为，Sn = 73.45%、Sp = 95.21%、ACC = 84.34%和 MCC = 0.704。比较其他 4 个模型，发现 FSID_PhSite 模型具有最高的识别总精度、特异性和马氏相关系数，总精度高出次优模型 CKSAAP_PhSite [12]约 6 个百分点。敏感性最高的模型是 DISPHOS [10]。

Table 2. Comparing the performance of different models in terms of serine (S) site identification on the independent dataset
表 2. 不同模型在独立检验集上对丝氨酸的识别性能比较

Models	ACC (%)	Sn (%)	Sp (%)	MCC
NetPhos [9]	64.91	78.90	55.64	0.343
DISPHOS [10]	70.10	81.03	62.86	0.432
PPRED [11]	62.87	72.62	56.42	0.286
CKSAAP_PhSite [12]	78.59	79.45	78.03	0.566
FSID_PhSite	84.34	73.45	95.21	0.704

注：NetPhos: artificial neural networks method for predicts phosphorylation sites; DISPHOS: disorder-enhanced phosphorylation predictor; PPRED: phosphorylation predictor; CKSAAP_PhSite: the composition of k-spaced amino acid apirs for prediction of protein phosphorylation sites; FSID_PhSite: feature selection based on increment of diversity for prediction of protein phosphorylation sites.

Table 3. Comparing the performance of different models in terms of threonine (T) site identification on the independent dataset
表 3. 不同模型在独立检验集上对苏氨酸的识别性能比较

Models	ACC (%)	Sn (%)	Sp (%)	MCC
NetPhos [9]	64.70	47.78	74.75	0.231
DISPHOS [10]	71.93	70.06	73.04	0.421
PPRED [11]	62.12	48.26	70.34	0.187
CKSAAP_PhSite [12]	78.98	79.16	78.88	0.567
FSID_PhSite	82.32	65.97	98.66	0.684

Table 4. Comparing the performance of different models in terms of tyrosine (Y) site identification on the independent dataset
表 4. 不同模型在独立检验集上对酪氨酸的识别性能比较

Models	ACC (%)	Sn (%)	Sp (%)	MCC
NetPhos [9]	59.92	45.80	69.30	0.154
DISPHOS [10]	66.62	55.24	74.19	0.298
PPRED [11]	56.42	43.01	65.35	0.084
CKSAAP_PhSite [12]	68.58	52.10	79.53	0.329
FSID_PhSite	68.89	47.18	90.59	0.420

由表 3 可见，我们的模型 FSID_PhSite 对磷酸化苏氨酸位点识别的敏感性为 65.97%、特异性为 98.66%、总精度为 82.32%和马氏相关系数为 0.684。除敏感性外，其他性能也显著优于其他模型。

由表 4 可见，我们的模型 FSID_PhSite 对磷酸化酪氨酸位点识别的敏感性为 47.18%、特异性为 90.59%和总精度为 68.89%和马氏相关系数为 0.420。比较其他模型，酪氨酸位点的识别结果类似于丝氨酸和苏氨酸。

由表 2、表 3 和表 4 我们还可以发现一个显著的特点，就是 FSID_PhSite 模型的特异性值都很高，三类磷酸化位点的特异性值均超过 90%。这表明 FSID_PhSite 模型对非磷酸化位点的识别准确度很高，这是其他四个模型均不具备的。由于非磷酸化位点数远大于磷酸化位点数，因此高特异性模型在推广性方面显然具备明显的优势。

4. 总结

我们提出一种精确的激酶无关磷酸化位点识别模型 FSID_PhSite，模型的主要特点是采用 FSID 方法进行高效特征选择，在独立测试集上的检验结果表明，FSID_PhSite 模型在磷酸化位点识别问题中具有优越的性能。FSID_PhSite 模型发现了磷酸化过程的一个可能的容错机制，并且对非磷酸化位点的识别具有极高的准确度。

致　谢

感谢内蒙古自治区自然科学基金资助项目(批准号：2015MS0331 和 2016MS0306)对本论文的支持。

参考文献

[1] Pinna, L.A. and Ruzzene, M. (1996) How do Protein Kinases Recognize Their Substrates? *Biochimica et Biophysica Acta*, **1314**, 191-225. https://doi.org/10.1016/S0167-4889(96)00083-3

[2] Wong, Y.H., Lee, T.Y., Liang, H.K., Huang, C.M., Wang, T.Y., Yang, Y.H., Chu, C.H., Huang, H.D., Ko, M.T. and Hwang, J.K. (2007) KinasePhos 2.0: A Web Server for Identifying Protein Kinase Specific Phosphorylation Sites Based on Sequences and Coupling Patterns. *Nucleic Acids Research*, **35**, W588-W594. https://doi.org/10.1093/nar/gkm322

[3] 张颖, 罗辽复, 吕军. 使用多样性增量预测磷酸化位点[J]. 内蒙古大学学报, 2008, 39(1): 34-39.

[4] Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L. and Yao, X. (2008) GPS 2.0, a Tool to Predict Kinase-Specific Phosphorylation Sites in Hierarchy. *Molecular and Cellular Proteomics*, **7**, 1598-1608. https://doi.org/10.1074/mcp.M700574-MCP200

[5] 白海燕, 吕军, 张颖, 等. 蛋白质磷酸化位点的识别[J]. 内蒙古工业大学学报, 2011, 30(2): 108-115.

[6] Trost, B., Kusalik, A. and Napper, S. (2016) Computational Analysis of the Predicted Evolutionary Conservation of Human Phosphorylation Sites. *PLoS One*, **11**, e0152809. https://doi.org/10.1371/journal.pone.0152809

[7] Karabulut, N.P. and Frishman, D. (2016) Sequence- and Structure-Based Analysis of Tissue-Specific Phosphorylation Sites. *PLoS One*, **11**, e0157896. https://doi.org/10.1371/journal.pone.0157896

[8] Zhao, Y.W., Lai, H.Y., Tang, H., Chen, W. and Lin, H. (2016) Prediction of Phosphothreonine Sites in Human Proteins by Fusing Different Features. *Scientific Reports*, **6**, 34817. https://doi.org/10.1038/srep34817

[9] Blom, N., Gammetltoft, S. and Brunak, S. (1999) Sequence and Structure-Based Prediction of Eukaryotic Protein Phosphorylation Sites. *Journal of Molecular Biology*, **294**, 1351-1362. https://doi.org/10.1006/jmbi.1999.3310

[10] Lakoucheva, L., Radivojac, P., Brown, C., *et al.* (2004) The Importance of Intrinsic Disorder for Protein Phosphorylation. *Nucleic Acids Research*, **32**, 1037. https://doi.org/10.1093/nar/gkh253

[11] Biswas, A.K., Noman, N. and Sikder, A.R. (2010) Machine Learning Approach to Predict Protein Phosphorylation Sites by Incorporating Evolutionary Information. *BMC Bioinformatics*, **11**, 273. https://doi.org/10.1186/1471-2105-11-273

[12] Zhao, X., Zhang, W., Xu, X., Ma, Z. and Yin, M. (2012) Prediction of Protein Phosphorylation Sites by Using the Composition of k-Spaced Amino Acid Pairs. *PLoS One*, **7**, e46302. https://doi.org/10.1371/journal.pone.0046302

[13] Chaudhuri, R. and Yang, J.Y. (2017) Cross-Species PTM Mapping from Phosphoproteomic Data. *Methods in Molecular Biology*, **1558**, 459-469. https://doi.org/10.1007/978-1-4939-6783-4_22

[14] Audagnotto, M. and Dal Peraro, M. (2017) Protein Post-Translational Modifications: *In Silico* Prediction Tools and Molecular Modeling. *Computational and Structural Biotechnology Journal*, **15**, 307-319. https://doi.org/10.1016/j.csbj.2017.03.004

[15] Drotár, P., Gazda, J. and Smékal, Z. (2015) An Experimental Comparison of Feature Selection Methods on Two-Class Biomedical Datasets. *Computers in Biology and Medicine*, **66**, 1-10. https://doi.org/10.1016/j.compbiomed.2015.08.010

[16] Yu, L. and Liu, H. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Fawcett, T. and Mishra, N., Eds., *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, The AAAI Press, Palo Alto, 856-863.

[17] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[18] Zou, Q., Zeng, J., Cao, L. and Ji, R. (2016) A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing*, **173**, 346-354. https://doi.org/10.1016/j.neucom.2014.12.123

[19] Yang, S.Q., Hu, S.S., Zhang, Y. and Lv, J. (2017) Application of Feature Selection Technology Based on Incremental of Diversity in Prediction of Flexible regions from Protein Sequences. *Letters in Organic Chemistry*, **14**, 621-624. https://doi.org/10.2174/1570178614666170221145333

[20] Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N. and Gibson, T.J. (2004) Phospho.ELM: A Database of Experimentally Verified Phosphorylation Sites in Eukaryotic Proteins. *BMC Bioinformatics*, **5**, 79. https://doi.org/10.1186/1471-2105-5-79

[21] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199