DOI:10.16515/j.cnki.32-1722/n.2020.02.009

细胞核蛋白质磷酸化位点的预测方法研究

余中洲[1],高 强[2],阴玉涵[2],张长青[3*],王 进[1*]

(1.南京大学生命科学学院,江苏 南京 210023;2.南京大学软件学院,江苏 南京 210023;
3.金陵科技学院研究生处,江苏 南京 211169)

摘 要:以包含细胞核蛋白质磷酸化位点的肽段为对象,将文本向量化方法与机器分类算法相组合,开展了蛋白质磷酸化位点特征筛选和组合模型评价。结果表明:磷酸化位点下游第一个氨基酸及磷酸化位点氨基酸种类是两个重要特征。独热编码与支持向量机的组合模型综合效果最好,它在训练集上的准确率达91.6%,精准度达94.0%,召回率达89.2%。测试集上的结果表明:该模型也显示出了良好的泛化能力。本模型为亚细胞层面磷酸化调控的精细分析提供了有效方法。

关键词:磷酸化位点;亚细胞定位;细胞核;激酶家族;预测模型

中图分类号:Q51 **文献标识码**:A **文章编号**:1672-755X(2020)02-0047-05

Study on the Prediction Method of Phosphorylation Sites of Cell Nucleus Proteins

YU Zhong-zhou[1], GAO Qiang[1], YIN Yu-han[1], ZHANG Chang-qing[2*], WANG Jin[1*]

(1. Nanjing University, Nanjing 210023, China; 2. Jinling Institute of Technology, Nanjing 211169, China)

Abstract: In this paper, the text vectorization methods were combined with machine classification algorithms to screen the features of protein phosphorylation site. And the combination models were evaluated for the prediction ability. The result showed that, the first amino acid downstream of phosphorylation site and the type of amino acid at phosphorylation site are two important features. The model of one-hot coding and svm combination showed the best performance with an accuracy of 91.6%, precision of 94.0% and 89.2% recall rate. The model also showed good generalization ability on the test dataset. This prediction model provides an effective approach for the in-depth analysis of phosphorylation regulation at subcellular level.

Key words: phosphorylation site; subcellular localization; nucleus; kinase family; prediction model

翻译后修饰在产生蛋白质结构与功能多样化方面起着重要作用。常见的蛋白质翻译后修饰包括磷酸化、乙酰化、泛素化等多种类型。研究发现,生物体中有200多种翻译后修饰[1],其中,磷酸化是最普遍、研究最多的蛋白质翻译后修饰。磷酸化修饰通过对底物蛋白质的丝氨酸、苏氨酸和酪氨酸上转移磷酸基团调节蛋白质活性和功能,从而调控各种细胞活动,包括增殖、分化和凋亡等[2],异常的磷酸化状态常与各种疾病包括癌症相关[3]。在蛋白质磷酸化研究中,磷酸化位点的确定最为关键,是认识蛋白质作用分子机理的基础。

收稿日期:2020-05-10
基金项目:江苏现代农业产业技术体系建设专项资金资助(JATS[2019]019)
作者简介:余中洲(1995—),男,河南信阳人,硕士研究生,主要从事生物信息学方面的研究。
通信作者:张长青(1974—),男,山西运城人,教授,博士,主要从事园艺生物信息学研究。
王进(1963—),女,江苏盐城人,教授,博士,主要从事生物信息学研究。

高通量质谱技术的广泛应用[4]使得蛋白质组磷酸化修饰位点的数据量得到了巨大增长，同时，通过计算的方法对磷酸化位点进行预测也日益受到关注。目前，磷酸化位点预测模型主要分为两类：一类是对磷酸化位点的不同类型训练相应模型，如 Ismail 等使用随机森林算法的 RF-Phos 模型[5]；另一类是根据磷酸化激酶的不同，构建激酶家族特异性的预测模型，如 Song 等创建的 Phosphopredict 模型[6]。另外 Luo 等使用深度学习算法创建的 DeepPhos 模型[7]和 Wang 等提出的 CapsNet 模型[8]，针对三种磷酸化位点和激酶家族特异性磷酸化位点，都训练了相应的模型。

细胞中的蛋白质磷酸化不是随机发生的，而是在各个细胞器中呈现多样的动态过程。这种空间分隔不仅能有效地调控蛋白质磷酸化，也能控制其它翻译后修饰的发生[9]。细胞内的蛋白质大多在特定的细胞器中进行磷酸化和去磷酸化，如 STAT1 与细胞膜上的受体相连，在信号分子与受体结合后被 Jak 酪氨酸激酶磷酸化，就会转移到细胞核中调控基因的转录[10]。因此，虽然大规模蛋白质组研究中检测了大量的磷酸化蛋白，但是，想要了解蛋白质磷酸化在特定细胞器中的功能以及如何发生作用的分子机制，还必须进一步获得磷酸化蛋白质的亚细胞定位[9]及其相应的磷酸化修饰图谱。为此，需要建立新的方法用于亚细胞水平的磷酸化位点研究，将蛋白质磷酸化分析的精度提高到亚细胞水平，从而深入理解其生物学功能。

1 材料与方法

1.1 数据收集与预处理

本文使用了来自两个数据库的数据，PhosphoSitePlus 数据库[11]和 SUBPHOS 数据库[12]，前者涵盖了迄今最全的激酶位点特异性数据，后者提供了磷酸化蛋白在细胞中 20 个细胞器的定位信息。我们从 SUBPHOS 数据库中提取出专门定位于细胞核中的磷酸化底物蛋白数据 119 个，从 PhosphoSitePlus 数据库提取含有 CDK 激酶(Cyclin-dependent kinases)[13-14]对应关系的磷酸化位点 381 个。使用 Cd-hit 以 70%的相似度为阈值对蛋白质进行去冗余处理[15]。去冗余后的 117 条底物蛋白多肽链上的 378 个阳性磷酸化位点数据构成了正数据集。在正数据集中抽取 20%的数据构建测试集，剩下的 80%正数据用于训练。负数据集由 117 条磷酸化底物蛋白上非磷酸化位点的丝氨酸、苏氨酸和酪氨酸随机抽样构成，其中训练集和测试集的样本大小与正数据集相同。

1.2 特征提取

以磷酸化位点及其上下游各 6 个氨基酸残基的肽段为对象，共提取出 417 个特征，具体包括：1)13 个位置的氨基酸残基符。每个位置包括 20 种常见氨基酸的单字符和 1 种替代符，产生共 21 种特征[16]。2)13 个位置的氨基酸残基所处二级结构。由预测工具[17]获得，每个位置涉及 H(螺旋)、C(卷曲)和 E(折叠)、X(未知)四种特征。3)13 个位置的氨基酸残基疏水性。由远程工具[17]计算获得，每个位置涉及 E(亲水)、B(疏水)、M(二者之间)、X(未知)四类特征。4)13 个位置的氨基酸残基有序性。由预测工具[17]预测获得，每个位置涉及有序、无序、未知三种特征。5)该肽段 10 个近邻中的正数据肽段占比。该参数首先由公式(1)计算某一肽段(称为肽段 A)与训练集中的每个肽段(称为肽段 B)的距离 $Dist(A,B)$[18]，然后取与肽段 A 距离最近的 10 个肽段得到肽段 A 的 10 个近邻序列，最后计算所含正数据集中肽段的百分比。

$$Dist(A,B) = 1 - \frac{\sum_{i=1}^{m+n+1} Sim(A[i],B[i])}{m+n+1} \tag{1}$$

$$Sim(a,b) = \frac{Matrix(a,b) - \min(Matrix)}{\max(Matrix) - \min(Matrix)} \tag{2}$$

式中，m 和 n 表示在磷酸化位点上下游的氨基酸数目，本研究中 m 和 n 均等于 6。a，b 表示肽段 A 和 B 同一位置上对应的氨基酸残基，$Matrix$ 表示 BLOSUM62 替代矩阵；max 和 min 分别表示矩阵中的最大值和最小值。

1.3　特征选择

机器学习中，处理过大特征集会降低机器学习速度，消耗过多计算资源，特定情况下还会降低准确率。为提高算法的效率和准确率，本研究使用 mRMR(maximum Relevance Minimum Redundancy)算法[19]对有效特征进行选择。

1.4　算法组合与评价

以独热编码和 Word2vec[20-21]方法对文本进行向量化处理，然后分别与支持向量机、随机森林、神经网络等分类算法组合，产生组合模型。以准确率、精准度、召回率、受试者工作特征曲线(Receiver Operating Characteristic curve，ROC)和 F_1 分值五个参数评价组合模型效率，涉及的计算公式如下：

$$\text{准确率}=\frac{TP+TN}{TP+TN+FP+FN};\text{精准度}=\frac{TP}{TP+FP};\text{召回率}=\frac{TP}{TP+FN};F_1\text{ 分值}=2\times\frac{TP}{2TP+FP+FN}$$

其中，TP 表示真阳性数目，TN 为真阴性数目，FP 为假阳性数，FN 为假阴性数。

2　结果与分析

2.1　模型评估

训练集数据共包含 304 条正数据和 304 条负数据，每条数据的特征维度是 417 维。使用特征选择工具 mRMR 算法对有效特征选择后，获得排名前 50 的特征(表 1)。从表 1 可见，50 个特征值共由 23 个不同的特征名称组成，包括 8 个不同位置的氨基酸残基，8 个不同位置氨基酸残基的二级结构，5 个不同位置氨基酸残基的疏水性，1 个氨基酸残基的有序性和磷酸化肽段的 K 近邻比率。其中，最重要的识别特征是氨基酸序列，尤其是酸化位点下游的第一个氨基酸和磷酸化位点的氨基酸种类，另外，K 近邻比率也有一定贡献。

表 1　CDK 激酶底物磷酸化位点的前 50 个特征信息

序号	特征索引	特征名称	序号	特征索引	特征名称
1	159	+1 AA	26	328	−6 solvent
2	145	0 AA	27	164	+1 AA
3	354	+1 solvent	28	298	0 structure
4	416	KNN	29	243	+5 AA
5	197	+2 AA	30	408	+4 order
6	297	0 structure	31	234	+5 AA
7	152	+1 AA	32	179	+2 AA
8	162	+1 AA	33	322	+6 structure
9	156	+1 AA	34	282	−4 structure
10	161	+1 AA	35	213	+4 AA
11	286	−3 structure	36	362	+3 solvent
12	203	+3 AA	37	155	+1 AA
13	299	0 structure	38	194	+3 AA
14	147	+1 structure	39	347	−1 solvent
15	374	+6 solvent	40	158	+1 AA
16	356	+1 solvent	41	56	−4 AA
17	150	+1 AA	42	200	+3 AA
18	96	−2 AA	43	293	−1 structure
19	301	+1 structure	44	86	−2 AA
20	148	+1 AA	45	51	−4 AA
21	355	+1 solvent	46	242	+5 AA
22	224	+4 AA	47	290	−2 structure
23	141	0 AA	48	163	+1 AA
24	303	+1 structure	49	307	+2 structure
25	149	+1 AA	50	165	+1 AA

注：重要性表示在 mRMR 进行特征筛选后的排名；特征索引表示特征在 417 维度中的索引；特征名称表示位置和磷酸化位点特征，其中 AA 表示氨基酸残基，structure 表示二级结构，solvent 表示疏水性，order 表示有序性，KNN 表示 K 近邻信息。

计算 CDK 激酶磷酸化肽段每个位置上的氨基酸频率并生成肽段序列模序图(图 1)。从图 1 可见,磷酸化位点下游的第一位氨基酸残基对 P 具有极强的偏向性。这与表 1 中的特征筛选结果,即+1 位氨基酸是最重要的识别特征,相吻合。

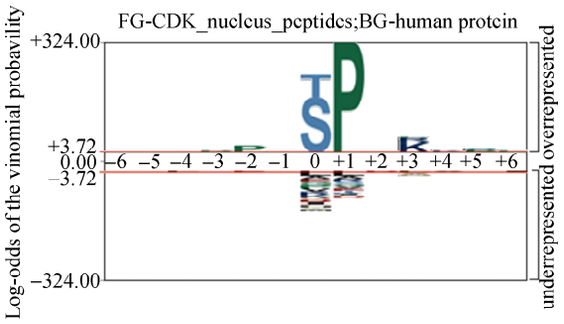

图 1 细胞核内 CDK 激酶磷酸化肽段的序列模序图

对训练集中的特征值进行筛选后,使用不同算法对训练集进行训练,每种算法重复训练 10 次,取平均值代表模型效果并进行评价。结果(表 2)表明,独热编码组合模型效果均优于 Word2vec 组合模型。在准确率、精准度以及 F_1 分值上,“独热编码+随机森林”组合效果最好,三个指标分别达到了 91.7%、94.3%和 91.4%;在召回率方面,“独热编码+神经网络”组合效果最好,达到了 89.8%;在受试者 ROC 特征曲线下面积方面,“独热编码+支持向量机”的组合最高,为 95.3%,同时这一组合的其他指标也排名靠前,因此以该模型当作最佳模型。

表 2 不同算法和文本向量化方法组合的性能指标

%

算法组合	准确率	精准度	召回率	F_1 分值	ROC 曲线下面积
独热编码+支持向量机	91.6	94.0	89.2	91.3	95.3
Word2vec+支持向量机	86.5	85.9	88.2	86.7	92.1
独热编码+随机森林	91.7	94.3	89.1	91.4	94.7
Word2vec+随机森林	90.1	92.2	88.1	89.8	93.6
独热编码+神经网络	89.3	89.2	89.8	89.3	95.2
Word2vec+神经网络	87.9	87.6	88.8	88.0	93.8

2.2 模型的泛化能力

使用测试集测试“独热编码+支持向量机”模型的泛化能力,测试集共包含 74 条正数据和 74 条负数据,同样每条数据的特征维度也是 417 维,选择排名前 50 的特征作为特征集。结果表明,“独热编码+支持向量机”模型的 ROC 的曲线下面积占 95.3%,其他各项指标中,召回率达 89.2%,F_1 为 91.3%,准确率 91.6%,精准度也达到了 94.0%。这表明了“独热编码+支持向量机”模型具有良好的泛化能力。

3 结论与讨论

本文以包含细胞核 CDK 激酶底物磷酸化位点的肽段为对象,通过对 6 个模型组合进行评价,构建了一个具有良好泛化能力的亚细胞层面激酶特异性磷酸化位点预测模型。与细胞层面的磷酸化预测模型相比,该模型考虑了磷酸化蛋白的亚细胞定位信息,这对于深入理解磷酸化在信号通路中的分子作用非常重要;与 Chen 等[12]提出的在亚细胞层面构建的磷酸化位点特异性模型相比,激酶家族特异性的模型更有利于阐释磷酸化的生物学功能和发现新的调控途径。另外,与深度学习模型[22]相比,该模型算法更具有解释性,例如,通过特征集合可以知道对于 CDK 激酶家族,磷酸化识别最重要的信息是磷酸化位点下游的第一位氨基酸残基。

“独热编码＋支持向量机”模型在训练集上准确率达到 91.6%。但随着亚细胞蛋白质磷酸化组学的进一步发展，增加训练集的数据量，以及采用更适合的分类算法，有可能进一步提高蛋白质磷酸化位点预测模型的预测能力。

参考文献：

[1] Duan G，Walther D. The roles of post-translational modifications in the context of protein interaction networks[J]. PLoS Comput Biol，2015，11(2)：e1004049－e1004052

[2] Singh V，Ram M，Kumar R，et al. Phosphorylation：Implications in cancer[J]. The Protein Journal，2017，36(1)：1－6

[3] Fleuren E D G，Zhang L，Wu J，et al. The kinome ‘at large’ in cancer[J]. Nature Reviews Cancer，2016，16(2)：83－98

[4] Beausoleil S A，Villen J，Gerber S A，et al. A probability-based approach for high-throughput protein phosphorylation analysis and site localization[J]. Nat Biotechnol，2006，24(10)：1285－1292

[5] Ismail H D，Jones A，Kim J H，et al. A novel general phosphorylation site prediction tool based on random forest[J]. Biomed Res Int，2016，166：3281590－3281596

[6] Song J，Wang H，Wang J，et al. PhosphoPredict：a bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection[J]. Sci Rep，2017，7(1)：6862－6869

[7] Luo F，Wang M，Liu Y，et al. DeepPhos：prediction of protein phosphorylation sites with deep learning[J]. Bioinformatics (Oxford，England)，2019，35(16)：2766－2773

[8] Wang D，Liang Y，Xu D. Capsule network for protein post-translational modification site prediction[J]. Bioinformatics，2019，35(14)：2386－2394

[9] Trost M，Bridon G，Desjardins M，et al. Subcellular phosphoproteomics[J]. Mass Spectrom Rev，2010，29(6)：962－990

[10] Xu X，Sarikas A，Dias-Santagata D C，et al. The CUL7 E3 ubiquitin ligase targets insulin receptor substrate 1 for ubiquitin-dependent degradation[J]. Mol Cell，2008，30(4)：403－414

[11] Hornbeck P V，Kornhauser J M，Tkachev S，et al. PhosphoSitePlus：a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse[J]. Nucleic Acids Res，2012，40：261－270

[12] Chen X，Shi S P，Suo S B，et al. Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity[J]. Bioinformatics，2015，31(2)：194－200

[13] Nurse P. Genetic control of cell size at cell division in yeast[J]. Nature，1975，256：547－551

[14] Weigel N L，Moore N L. Cyclins，cyclin dependent kinases，and regulation of steroid receptor action[J]. Mol Cell Endocrinol，2007(5)，265－266

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

[16] Li T，Du P，Xu N. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources[J]. PLoS One，2010，5(11)：e15411

[17] Wang S，Li W，Liu S，et al. RaptorX-Property：a web server for protein structure property prediction[J]. Nucleic Acids Res，2016，44：430－435

[18] Chen Z，Zhao P，Li F，et al. IFeature：a python package and web server for features extraction and selection from protein and peptide sequences[J]. Bioinformatics，2018，34(14)：2499－2502

[19] Peng H，Long F，Ding C. Feature selection based on mutual information：criteria of max-dependency，max-relevance，and min-redundancy[J]. IEEE Trans Pattern Anal Mach Intell，2005，27(8)：1226－1238

[20] Le Q，Mikolov T. Distributed representations of sentences and documents[J]. 31st International Conference on Machine Learning，2014(4)：125－132

[21] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems，2013，26：369－376

[22] Luo F，Wang M，Liu Y，et al. DeepPhos：prediction of protein phosphorylation sites with deep learning[J]. Bioinformatics，2019，35(16)：2766－2773

（责任编辑：谭彩霞）