PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(7): 691~698
www.pibb.ac.cn

基于三类特征融合的 O-糖基化位点预测 *

向 妍[1)**] 陈 渊[1)**] 谭泗桥[2)***] 袁哲明[1, 3)***]

([1)] 湖南农业大学植物病虫害生物学与防控湖南省重点实验室，长沙 410128；[2)] 湖南农业大学信息科学技术学院，长沙 410128；
[3)] 湖南农业大学，湖南省作物种质创新与资源利用重点实验室， 长沙 410128)

摘要 糖基化是蛋白质翻译后的主要修饰，O- 糖基化的固定模式未知，高精度识别 O- 糖基化位点是机器学习面临的挑战性问题．以迄今最大的人 O- 糖基化位点 Steentoft 数据集为基础，本文首次提出了基于位置的卡方差表特征 χ^2-pos，融合伪氨基酸序列进化信息 PsePSSM 以及无方向的 k 间隔氨基酸对组分 Undirected-CKSAAP 表征序列，构建 5 个正负样本均衡的支持向量机分类器，经加权投票，独立测试准确率、Matthew 相关系数及 ROC 曲线下面积，分别达到了 89.62%、0.79、0.96，明显优于文献报道结果．χ^2-pos、PsePSSM 与 Undirected-CKSAAP 三种特征的融合在蛋白质糖基化、磷酸化等位点预测中有广泛应用前景．

关键词 O- 糖基化位点预测，卡方差表特征，伪氨基酸序列进化信息，无方向的 k 间隔氨基酸对组分，加权投票
学科分类号 Q51，Q61 **DOI**: 10.16476/j.pibb.2016.0002

蛋白质在生命体内存在糖基化、磷酸化、泛素化、甲基化、脂基化和乙酰化等多种翻译后修饰，其中糖基化是最主要的修饰之一，50%以上的蛋白质存在糖基化现象[1]，涉及细胞免疫、蛋白质翻译调控、蛋白质降解等众多生物学过程[2-4]．蛋白质糖基化主要包括 O- 糖基化、N- 糖基化与 C- 糖基化[5]， N- 糖基化与 C- 糖基化含有固定模式，而 O- 糖基化位点 S/T 及其侧翼序列缺乏保守性[5-8]．实验识别 O- 糖基化位点耗时费力，如何从包含大量虚假位点的 S/T 残基中有效识别 O- 糖基化位点是机器学习面临的挑战性问题[9-11]．

现有 O- 糖基化位点识别研究中，训练集正负样本比例有 89∶126[12-13]、209∶109[14]、261∶1305[15]、328∶285[16]，测试集正负样本数有 26∶22[14]、82∶117[12]等，总体来说样本偏少．窗口长度多经验性地选取 9、15、23 或 41 个残基．序列表征方法有基于位置的残基 01 编码与氨基酸理化性质[15]、基于组分的 20 种氨基酸频率[12]、基于组分与关联的 k 间隔氨基酸对组分(composition of k-spaced amino acid pairs，CKSAAP)[16]、预测的二级结构[17]、GO 功能注释以及反映序列进化信息的位置特异型打分矩阵(position-specific scoring matrix，PSSM)等．特征选择方法常采用最大相关最小冗余与递归特征消减[13, 15-16]．分类器多采用支持向量机(supported vector machine，SVM)[12, 15-16, 18-19]、神经网络[13, 20]、随机森林[21]．独立测试或交叉测试准确率范围(accuracy，Ac)为 79.50%～85%，Matthew 相关系数范围(Matthew's correlation coefficient，MCC)为 0.58～0.67，预测精度尚有较大提升空间[12-13, 15-16]．

2013 年，Steentoft 等[17]以高通量实验鉴定了 662 条人蛋白质序列中的 O- 糖基化位点，包含 2 168 个真实位点(853S∶1315T)与 76 845 个虚假位点(43387S∶33458T)．该数据集是迄今为止报道的样本数最多的 O- 糖基化位点数据集．本文首次提出了基于位置的卡方差表法(χ^2-pos)，融合伪氨基酸序列进化信息(pseudo position specific scoring

* 高等学校博士学科点专项科研基金(20124320110002)，湖南省自然科学基金(14JJ2082)和长沙市科技计划项目(K1406018-21)资助．
** 共同第一作者．
*** 通讯联系人．
谭泗桥．Tel: 0731-84613956, E-mail: tsq@hunau.net
袁哲明．Tel: 0731-84613956, E-mail: zhmyuan@sina.com
收稿日期：2016-01-05，接受日期：2016-05-16

matrix，PsePSSM)以及无方向的 k 间隔氨基酸对组分 (undirected composition of k-spaced amino acid pairs，Undirected-CKSAAP)，采用相同正样本不同负样本构建了 5 个 1∶1 SVM 分类器并经加权投票，在该数据集上获得了较理想的独立预测精度，结果报道如下.

1　数据与方法

1.1　数据集

从 Steentoft 等[17]报道的 662 条人蛋白质序列中以 S/T 为中心、窗口长度 61 个残基截取样本序列，得正样本 2 005 个，负样本 69 881 个. 从正样本中随机抽取 1 403 个样本用作训练，命名为 Train-pos；剩余的 602 个样本用作独立测试，命名为 Test-pos.从负样本中随机无放回抽取 1 403 个样本用作训练，重复 5 次，依次命名为 Train-neg-1，Train-neg-2，…，Train-neg-5；在剩余负样本中再随机抽取 602 个用作独立测试，命名为 Test-neg. 则 5 套 1∶1 均衡训练集依次为 Train-pos∶Train-neg-1，Train-pos∶Train-neg-2，…，Train-pos∶Train-neg-5；一套 1∶5 不均衡训练集为 Train-pos∶(Train-neg-1+Train-neg-2+…+Train-neg-5)，简记为 Train-pos∶Train-neg；独立测试集为 Test-pos: Test-neg，共 1 204 个样本.

1.2　基于位置的卡方差表法(χ^2-pos)

以 Train-pos∶Train-neg-1 为例，统计 20 种氨基酸残基在第 i 个位置(i=1，2，…，61)正负样本中的频次可得如下 2×20 列联表(表 1).

Table 1　Frequency distribution of amino acid residues between positives and negatives for the ith position

Sample	Amino acid residue						Total
	1(A)	2(R)	…	j	…	20(V)	
True	$f_{i,1}^{+}$	$f_{i,2}^{+}$	…	$f_{i,j}^{+}$	…	$f_{i,20}^{+}$	f_i^{+}
False	$f_{i,1}^{-}$	$f_{i,2}^{-}$	…	$f_{i,j}^{-}$	…	$f_{i,20}^{-}$	f_i^{-}
Total	$f_{i,1}$	$f_{i,2}$	…	$f_{i,j}$	…	$f_{i,20}$	N

表中，$f_{i,j}^{+}$为第 j 种残基在第 i 个位置正样本中出现的频次，$f_{i,j}^{-}$为第 j 种残基在第 i 个位置负样本中出现的频次，$f_{i,j}$为第 j 种残基在第 i 个位置所有样本中出现的频次，f_i^{+}=f_i^{-}= 1 403，N = 2 806. 其卡方值按下式计算：

$$\chi^2=\frac{N^2}{f_i^{+}\times f_i^{-}}\left[\sum_{i=1}^{20}\frac{f_{i,j}^{+2}}{f_{i,j}}-\frac{f_i^{+2}}{N}\right] \quad (1)$$

若新增一训练样本，其第 i 个位置为第 j 种氨基酸；假设其为正样本，用 $f_{i,j}^{+}$+1 替换 $f_{i,j}^{+}$，按式 (1) 算得一个卡方值 $\chi_{i,j}^{+}$；再假设其为负样本，用 $f_{i,j}^{-}$+1替换 $f_{i,j}^{-}$，按式 (1) 算得一个卡方值 $\chi_{i,j}^{-}$. 则第 i 个位置为残基的卡方差表得分为 $\Delta\chi_{i,j}=\chi_{i,j}^{+}-\chi_{i,j}^{-}$. 以窗口长度 61 的第一套训练集 (Train-pos: Train-neg-1)为例，可得如下 20×61 卡方差表(表 2).

从而，训练集和测试集中每条序列凡第 i 个位置出现第 j 种氨基酸，可赋值 $\Delta\chi_{i,j}$.

Table 2　Chi-square difference values for 20 amino acid residues and 61 positions

Amino acid residue	Protein(P)			
	Position(−30)	…	Position(i)	Position(+30)
1(A)	$\Delta\chi_{-30,1}$	…	$\Delta\chi_{i,1}$	$\Delta\chi_{30,1}$
…	…	…	…	…
j	$\Delta\chi_{-30,j}$	…	$\Delta\chi_{i,j}$	$\Delta\chi_{30,j}$
20(V)	$\Delta\chi_{-30,20}$	…	$\Delta\chi_{i,20}$	$\Delta\chi_{30,20}$

1.3 伪氨基酸序列进化信息(PsePSSM)

反映序列进化信息的 PSSM 由本地化的 PSI-BLAST[22]程序搜索 Swiss-Prot 数据库获得，迭代 3 次，*E*-value 值设置为 0.001．PSI-BLAST 是 BLAST 的扩展，它允许迭代搜索，并擅于发现序列间远缘关系[22]．长为 L 的一条序列其规格化 PSSM 矩阵如下：

$$P_{\text{pssm}}=P_{i-j}(i=1, 2, \cdots, L;\ j=1, 2, \cdots, 20) \quad (2)$$

式中，P_{i-j} 为第 i 位残基突变为第 j 种天然氨基酸的归一化得分值．

为了解决序列不等长导致的样本特征维数不一致的问题，Shen 等[23]提出了伪氨基酸序列进化信息 PsePSSM：

$$\begin{cases} P^{m}_{\text{psePSSM}}=[G^{m}_{1}, G^{m}_{2}, \cdots, G^{m}_{j}, \cdots, G^{m}_{20}] \\ G^{m}_{j}=\dfrac{1}{L-m}\displaystyle\sum_{i=1}^{L-m}\left[P_{i-j}-P_{(i+m)-j}\right]^{2} \end{cases} (i=1, 2, \cdots, L;\ j=1, 2, \cdots, 20;\ m=0, 1, \cdots, \lambda) \quad (3)$$

式中，G^{m}_{j} 是第 j 种氨基酸在间隔距离为 m 时的相关因子；$0\leqslant\lambda<L$，λ 为最大间隔距离．这样，每条序列可得 $20\times(\lambda+1)$维的 PsePSSM 特征．

1.4 无方向的 k 间隔氨基酸对组分(Undirected-CKSAAP)

k 间隔氨基酸对组分 CKSAAP 是序列中由 k 个残基分隔的氨基酸对出现的频率[16]．每条序列可得 $400\times(k+1)$维的 CKSAAP 特征．CKSAAP 兼顾了序列的组分与上下文关联，且对不等长序列同样适用；但当 k 较大或序列较短时特征矩阵会过于稀疏，本文取 $k\leqslant3$．

在蛋白质空间结构中，AC 与 CA、A*C 与 C*A 是难以区分左右方向的，即 AC 与 CA、A*C 与 C*A 可视为相同项并合并[24]．本文据此提出了无方向的 k 间隔氨基酸对组分 Undirected-CKSAAP，一定程度缓解了矩阵稀疏的问题．每条序列可得 $210\times(k+1)$维 Undirected-CKSAAP 特征．

1.5 SVM 分类器与加权投票

分类器采用 Libsvm 3.1[25]，核函数固定为径向基核，参数 c、g 基于训练集由 grid.py 经 5-fold 交叉测试搜索自动获取．SVM 对特征维数不敏感，且研究中发现，常用的特征选择方法最大相关最小冗余、递归特征消减在该数据集上作用甚微，甚至有不利影响，故本文未进行特征选择．

为解决正负样本不均衡的问题并降低样本数较大时 SVM 的训练耗时(SVM 的计算复杂度为 $\Phi(N^3)$，N 为样本数)，本文构建了 5 个 1∶1 分类器．对第 i 个待测样本，设第 j 个分类器判定其属于正类的概率为 W_{ij}，属于负类的概率为 $1-W_{ij}$．加权投票策略为：若 $\sum_{j=1}^{5}W_{ij}>(1-\sum_{j=1}^{5}W_{ij})$，则第 i 个待测样本最终判为正类；否则判为负类．

1.6 模型评价指标

采用敏感性(sensitivity, Sn)、特异性(specificity, Sp)、准确率(accuracy, Ac)和 Matthew 相关系数 MCC 等指标评估模型表现：

$$Sn=\frac{TP}{TP+FN}\times100\%$$

$$Sp=\frac{TN}{TN+FP}\times100\%$$

$$Ac=\frac{TP+TN}{TP+TN+FP+FN}\times100\%$$

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

式中，TP(ture positive)、TN(ture negative)、FN(false negative)、FP(false positive)分别表示正样本判对数、负样本判对数、正样本判错数、负样本判错数．

接收者操作特征曲线(receive operating characteristic curve，ROC)也被广泛用于衡量预测模型性能[26-27]，它是以 $1-Sp$ 为横坐标，Sn 为纵坐标，表示在所有可能的阈值情况下所绘制的曲线．ROC 曲线下的面积(area under ROC curve，AUC)在 0～1 范围内，若预测性能越好，则 AUC 的值越接近于 1．

2 结果与分析

2.1 重要位置确定

以 Train-pos∶Train-neg-1 为例，各位置卡方得分如图 1．可见，除–1 位残基外，O- 糖基化位点侧翼序列保守性整体较弱，但整体仍呈现离中心点越远，保守性越弱的规律．

除中心位置 0 外，各位置按卡方得分从小到大排序，并以 χ^2-pos 表征，按序逐个剔除特征，训练集 5-fold 交叉测试精度变化如图 2．可见，剔除 –27、–30、–19、+13、+28、+27、+18、–24、–21、+26、–29 等 11 个位置后，交叉测试精度最

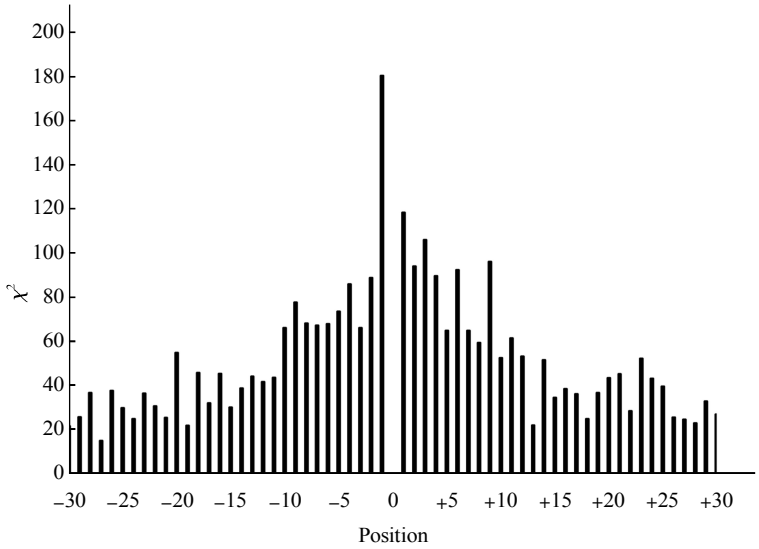

Fig. 1　Chi-square values for different positions in protein sequence

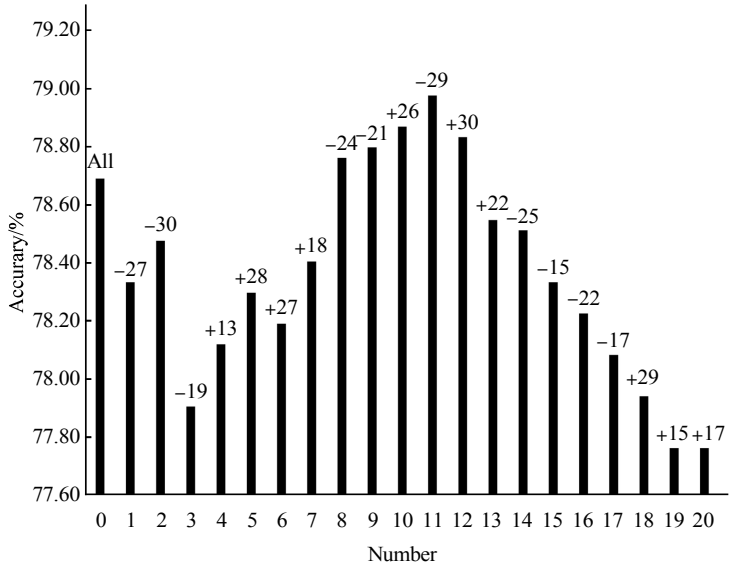

Fig. 2　The 5-fold cross accuracy of χ^2-pos features excluding one by one

高．最终保留 50 个重要位置(0 位保留用以区分中心 S/T 残基)，每条序列得 50 维χ^2-pos 特征．

2.2　PsePSSM 参数优化

PsePSSM 包含窗口长度与最大间隔距离 λ 两个待定参数．本文窗口长度最短取 23 个残基，最长取 59 个残基，左右等长，步长为 4 个残基，共 10 水平．最大间隔距离 λ 最小取 10，最长取 18，步长为 2 个残基，共 5 水平．全组合共 50 个处理．对每一组合处理，各序列以 PsePSSM 表征，训练集 5-fold 交叉测试精度如图 3．可见，当窗口长度为 47、最大间隔距离 λ 为 16 时交叉测试精度最高．因此，间隔距离 m 应取 0～16．这样，每条序列可得 20×(16+1)=340 维的 PsePSSM 特征．

2.3　Undirected-CKSAAP 的 k 值确定

基于 PsePSSM 优化的窗口长度 47，依次取 k=0、1、2、3，各序列以 Undirected-CKSAAP 表征，训练集 5-fold 交叉测试精度 Ac 为 81.93%、82.72%、82.93%、82.22%．故最优 k 值为 2，每条

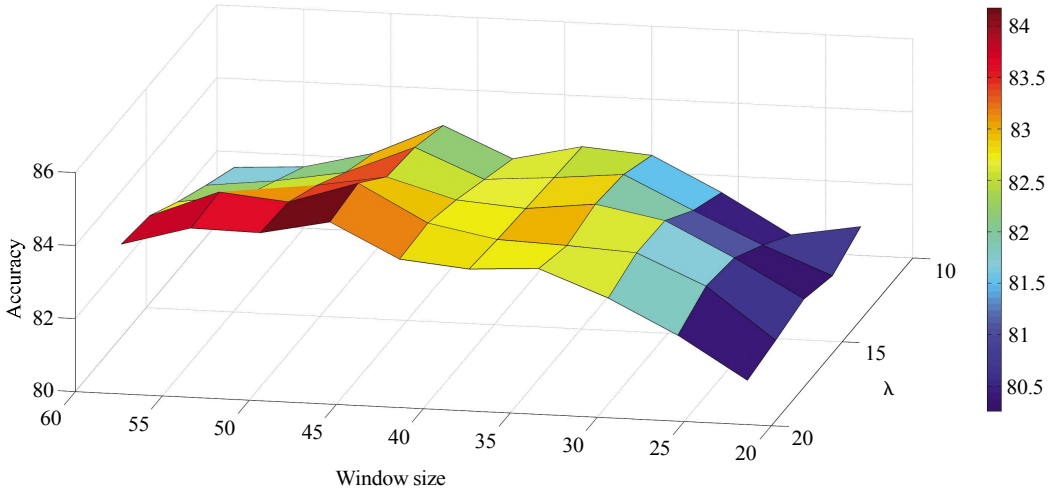

Fig. 3　The 5-fold cross accuracy for different window size coupled with the largest interval distances

序列可得 210×(2+1)=630 维 Undirected-CKSAAP 特征.

2.4　特征融合与加权投票结果

基于 50 个重要位置、最优窗口长度 47、最大间隔距离 λ=16、最优 k=2，以第一套训练集(Train-pos:Train-neg-1)为例，3 类特征单独、两两融合、全融合独立预测表现如表 3．可见，3 类特征全融合预测表现最优.

Table 3　The independent accuracy based on different encoding schemes

Feature	*Sn*/%	*Sp*/%	*Ac*/%	*MCC*
χ^2-pos	72.26	69.60	70.93	0.42
PsePSSM	82.06	86.54	84.30	0.69
Undirected-CKSAAP	79.40	90.53	84.97	0.70
χ^2-pos+PsePSSM	84.55	84.39	84.46	0.69
χ^2-pos+Undirected-CKSAAP	88.37	84.55	86.46	0.73
PsePSSM+Undirected-CKSAAP	86.85	85.22	86.05	0.72
χ^2-pos +PsePSSM +Undirected-CKSAAP	88.04	88.54	88.29	0.77

融合 3 类特征，5 个 1∶1 均衡子训练集加权投票法与单个 1∶5 不均衡训练集的独立预测结果见表 4．可见：a．加权投票 *Ac* 89.62%较 5 个 1∶1 均衡子训练集的平均 *Ac* 88.80%略有提高．b．Libsvm 的–*b* 参数可获得每个待测样本属于正样本的概率．以 P > 0.5 为标准，1∶5 不均衡训练集的独立预测 *Ac* 为 84.88%，较 1∶1 均衡子训练集明显下降，但若以训练集正样本先验概率即 P > 1/6 为标准，1∶5 不均衡训练集的独立预测 *Ac* 为 88.29%，与 1∶1 均衡子训练集接近．c．多个均衡子训练集加权投票法或先验概率校正法是解决训练集正负样本不平衡问题的有效手段，结合计算复杂度与预测稳健性，推荐使用多个均衡子训练集加权投票.

5 个均衡子训练集加权投票的 ROC 曲线见图 4, 其 AUC 值为 0.96.

就我们所知，该数据集迄今仅 Steentoft 等以 TMHMM 预测的跨膜区、NetSurfP 预测的表面可接触性、DISEMBL 预测的蛋白质无序区为特征报道过独立预测结果，其 *Ac*、*MCC* 分别为 83%、0.71[17]，本文加权投票法 *Ac*、*MCC* 分别为 89.62%、0.79，明显优于文献报道.

Table 4 The independent accuracy in the different train sets

Train set	*Sn*/%	*Sp*/%	*Ac*/%	*MCC*
Train-pos:Train-neg-1	88.04	88.54	88.29	0.77
Train-pos:Train-neg-2	88.87	88.87	88.87	0.78
Train-pos:Train-neg-3	88.37	90.70	89.53	0.79
Train-pos:Train-neg-4	88.87	89.04	88.95	0.78
Train-pos:Train-neg-5	89.20	87.54	88.37	0.77
Weighted voting	90.30	88.80	89.62	0.79
Train-pos:Train-neg (P > 0.5)	72.09	97.67	84.88	0.72
Train-pos:Train-neg (P > 1/6)	87.38	89.20	88.29	0.77

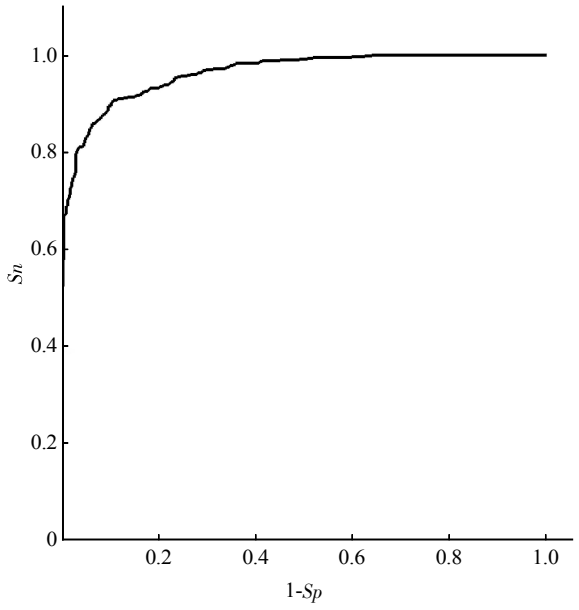

Fig. 4 ROC curves of weighted voting results

3 讨 论

3.1 χ^2-pos 特征的优点

基于位置的序列表征常见有残基 01 编码与氨基酸理化性质编码．以 Train-pos：Train-neg-1 为例，基于本文给定的 50 个重要位置，用 5-fold 交叉测试检验以下 4 种编码方式：残基 01 编码、531 种氨基酸理化性质编码、5 类氨基酸理化性质编码[28]及 χ^2-pos 编码(表 5)．结果表明，χ^2-pos 编码的特征维数最少，且预测精度最高，明显优于残基 01 编码及氨基酸理化性质编码．a．采用残基 01 编码，每个位置需用 20 维 0/1 特征表示，特征矩阵非常稀疏，且不能反映残基间某一理化性质的差异程度．例如某位点含 S、N、W 三种残基，假定因变量 Y 主要与该位点残基疏水性有关，其疏水性指数分别为 0.05、0.06、2.65；S-N 间疏水性相差极小，N-W 间疏水性相差较大，但按 01 编码，S-N 与 N-W 间汉明距离均为 1．b．AAindex 数据库(http://www.genome.jp/aaindex/)含 20 种天然氨基酸的 531 种理化性质．采用以上 531 种氨基酸理化性质编码时，因事先并不知道因变量 Y 与哪些理化性质相关，每个位置需用 531 维理化性质表示，其中存在大量无关特征与冗余特征．c．Atchley 等[28]对 531 种理化性质多元统计分析提出 5 类理化性质，分别是极性、二级结构、分子体积、密码子多样性及静电荷．采用以上 5 类理化性质编码时，蛋白质序列的特征维数大幅度降低，并已应用于 O-糖基化位点预测[12, 15]及其他领域[29]．d．但无论 531 种理化性质编码还是 5 类理化性质编码，其预测精度总是低于 χ^2-pos 编码的预测精度．

Table 5 The 5-fold cross accuracy based on different positional features

Feature	Feature number	*Ac*/%
Binary	1 000	70.38
Physicochemical properties(531)	26 550	71.06
Physicochemical properties(5)	250	69.00
χ^2-pos	50	78.97

当正负样本序列高度相似时(见下)，CKSAAP 等基于组分与关联的方法提取的特征向量将非常近似，导致分类器无法有效区分．相反，χ^2-pos 等基于位置的方法提取的特征向量将差异甚大，分类器能有效区分．

正样本 P1: RKSSNLDKDSSLSFQ**S**TQVPERR-HASLATVF

负样本 N1: KSSNLDKDSSLSFQS**T**QVPERRH-

ASLATVFP

χ^2-pos 是基于数据驱动的，能反映同一位置不同残基与不同位置同一残基的得分差异，能有效区分高度相似的正负样本序列，并具特征维数少、冗余度低、特征矩阵不稀疏等优点，在分子序列表征中可望有广泛应用前景.

3.2 改进 PSSM、CKSAAP 与三类特征融合的必要性

基于位置的 PSSM 尽管能有效区分高度相似的正负样本序列，但当存在插入、缺失时缺乏容错性．PsePSSM 不仅可解决序列不等长的问题，同时使得序列进化信息特征具容错能力．以 Train-pos：Train-neg-1 为例，取最优窗口长度 47、最优最大间隔距离 λ=16、PSSM 特征维数为 47×20=940、训练集 5-fold 交叉测试 Ac 为 71.42%；PsePSSM 特征维数为 $20\times(16+1)=340$、训练集 5-fold 交叉测试 Ac 为 84.18%．可见，PsePSSM 明显优于 PSSM.

仍以 Train-pos：Train-neg-1 为例，取最优窗口长度 47、最优 k=2，基于组分与关联的有方向 CKSAAP 特征维数为 $400\times(2+1)=1200$，训练集 5-fold 交叉测试 Ac 为 82.18%，无方向的 Undirected-CKSAAP 特征维数为 $210\times(2+1)=630$，训练集 5-fold 交叉测试 Ac 为 82.93%．可见，Undirected-CKSAAP 优于 CKSAAP.

基于位置的 χ^2-pos 能反映同一位置不同残基与不同位置同一残基的得分差异，能有效区分高度相似的正负样本序列，但对残基插入、缺失敏感．PsePSSM 能反映序列进化信息甚至是远缘关系，Undirected-CKSAAP 在反映序列组分的同时能捕捉序列的上下文关联，PsePSSM 与 Undirected-CKSAAP 均具容错性．三类特征从不同侧面对序列进行表征，互相补充，缺一不可，融合后预测表现最优(表 3).

参 考 文 献

[1] Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochimica et Biophysica Acta (BBA)-General Subjects, 1999, **1473**(1): 4−8

[2] Geoghegan K F, Song X, Hoth L R, *et al*. Unexpected mucin-type O-glycosylation and host-specific N-glycosylation of human recombinant interleukin-17A expressed in a human kidney cell line. Protein Expression and Purification, 2013, **87**(1): 27−34

[3] Gill D J, Chia J, Senewiratne J, *et al*. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. The Journal of Cell Biology, 2010, **189**(5): 843−858

[4] Katrine T B G S, Clausen H. Site-specific protein O-glycosylation modulates proprotein processing-deciphering specific functions of the large polypeptide GalNAc-transferase gene family. Biochimica et Biophysica Acta (BBA)-General Subjects, 2012, **1820** (12): 2079−2094

[5] Blom N, Sicheritz-Pontén T, Gupta R, *et al*. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 2004, **4**(6): 1633−1649

[6] Hart G W. Glycosylation. Current Opinion in Cell Biology, 1992, **4**(6): 1017−1023

[7] Wilson I B, Gavel Y, Von Heijne G. Amino acid distributions around O-linked glycosylation sites. Biochem. J, 1991, **275** (2): 529−534

[8] Christlet T H T, Veluraja K. Database analysis of O-glycosylation sites in proteins. Biophysical Journal, 2001, **80**(2): 952−960

[9] Julenius K, Mølgaard A, Gupta R, *et al*. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology, 2005, **15**(2): 153−164

[10] Haltiwanger R S, Lowe J B. Role of glycosylation in development. Annual Review of Biochemistry, 2004, **73**(1): 491−537

[11] Jensen O N. Interpreting the protein language using proteomics. Nature Reviews Molecular Cell Biology, 2006, **7**(6): 391−403

[12] Liu Y, Gu W, Zhang W, *et al*. Predict and analyze protein glycation sites with the mRMR and IFS methods. BioMed Research International, 2015, **2015**(2015): 1−6

[13] Johansen M B, Kiemer L, Brunak S. Analysis and prediction of mammalian protein glycation. Glycobiology, 2006, **16**(9): 844−853

[14] Niu B, Lu W, Ding J, *et al*. A two-stage method for O-glycosylation site prediction. Chemometrics and Intelligent Laboratory Systems, 2011, **108**(2): 142−145

[15] Li S, Liu B, Zeng R, *et al*. Predicting O-glycosylation sites in mammalian proteins by using SVMs. Computational Biology and Chemistry, 2006, **30**(3): 203−208

[16] Chen Y Z, Tang Y R, Sheng Z Y, *et al*. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. BMC Bioinformatics, 2008, **9**(1): 101

[17] Steentoft C, Vakhrushev S Y, Joshi H J, *et al*. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. The EMBO Journal, 2013, **32**(10): 1478−1488

[18] Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for predicting the specificity of GalNAc-transferase. Peptides, 2002, **23**(1): 205−208

[19]Caragea C, Sinapov J, Silvescu A, *et al*. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC Bioinformatics, 2007, **8**(1): 1−13

[20]Cai Y D, Yu H, Chou K C. Artificial neural network method for predicting the specificity of GalNAc-transferase. Journal of Protein Chemistry, 1997, **16**(7): 689−700

[21] Hamby S E, Hirst J D. Prediction of glycosylation sites using random forests. BMC Bioinformatics, 2008, **9**(1): 1−13

[22] Schäffer A A, Aravind L, Madden T L, *et al*. Improving the

accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research, 2001, **29**(14): 2994-3005

[23] Shen H B, Chou K C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering Design and Selection, 2007, **20**(11): 561-567

[24] Chou K C. Using pair-coupled amino acid composition to predict protein secondary structure content. Journal of Protein Chemistry, 1999, **18**(4): 473-480

[25] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, **2**(3): 1-27

[26] Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Critical Care, 2004, **8**(6): 508-512

[27] Centor R M. Signal detectability the use of ROC curves and their analyses. Medical Decision Making, 1991, **11**(2): 102-106

[28] Atchley W R, Zhao J, Fernandes A D, *et al*. Solving the protein sequence metric problem. Proc Natl Acad Sci USA, 2005, **102**(18): 6395-6400

[29] Zou C, Gong J, Li H. An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. BMC Bioinformatics, 2013, **14**(1): 1-14

Predicting O-glycosylation Sites by Combining Three Different Types of Features*

XIANG Yan[1)**], CHEN Yuan[1)**], TAN Si-Qiao[2)***], YUAN Zhe-Ming[1, 3)***]

([1)] *Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha* 410128, China;

[2)] *College of Information Science and Technology, Hunan Agricultural University, Changsha* 410128, China;

[3)] *Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha* 410128, China)

Abstract Glycosylation is a major modification process in post-translational modification of protein. Accurate prediction of O-linked glycosylation sites is a big challenging faced by machine-learning, for the fixed-model of O-linked glycosylation is not yet known. In this paper, on the basis of the largest-ever Steentoft database up to now, a new feature——chi-square score difference table method based on position (χ^2-pos) was first proposed, which combined with pseudo position-specific scoring matrix (PsePSSM) and undirected composition of k-spaced amino acid pairs (Undirected-CKSAAP) were used to present protein sequences. Then 5 support vector machines models were constructed with the same proportion of positive and negative samples. At last, by weighted voting, our results showed that the prediction accuracy, Matthew's correlation coefficient and area under ROC curve reached 89.62%, 0.79 and 0.96 respectively. They were superior to the literature report. It also demonstrated that the combination of three different features χ^2-pos, PsePSSM and Undirected-CKSAAP has extensive application prospect in protein sites prediction such as glycosylation and phosphorylation.

Key words O-glycosylation prediction, chi-square score difference table, pseudo position-specific scoring matrix, undirected composition of k-spaced amino acid pairs, weighted voting

DOI: 10.16476/j.pibb.2016.0002

*This work was supported by grants from Specialized Research Fund for the Doctoral Program of Higher Education (20124320110002), The Natural Science Foundation of Hunan Province, China (14JJ2082) and The Science and Technology Planning Projects of Changsha, China (K1406018-21).

**These authors contributed equally to this work.

***Corresponding author.

TAN Si-Qiao. Tel: +86-731-84613956, E-mail: tsq@hunau.net

YUAN Zhe-Ming. Tel: +86-731-84613956, E-mail: zhmyuan@sina.com

Received: January 5, 2016 Accepted: May 16, 2016