

基于特征选择和支持向量机的 HIV-1 型蛋白酶剪切位点预测

袁哲明^{1,2}, 张弘杨^{1,2}, 陈渊¹

(1. 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; 2. 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)

摘要:为了提高 HIV-1 型蛋白酶剪切位点的预测准确性,提出一种基于特征选择和支持向量机的剪切位点预测模型。首先,通过对 5830 个样本的 HIV-1 型蛋白酶剪切位点数据集进行分析,根据最小冗余最大相关理念,采用可自动终止法选择剪切位点的特征向量;然后,将特征向量输入到支持向量机进行学习和训练,建立 HIV-1 型蛋白酶剪切位点的分类模型;最后,采用 Matlab 2014 的仿真工具箱进行仿真测试。实验结果表明:本文模型在特征最少的条件下,剪切位点预测精度优于参比模型及文献报道,且所选择的特征向量具有较好的可解释性及生物学意义。

关键词:生物物理学; 剪切位点预测; 特征选择; 最小冗余最大相关; 支持向量机

中图分类号:Q6 **文献标志码:**A **文章编号:**1671-5497(2017)02-0639-08

DOI:10.13229/j.cnki.jdxbgxb201702040

HIV-1 protease cleavage site prediction based on feature selection and support vector machine

YUAN Zhe-ming^{1,2}, ZHANG Hong-yang^{1,2}, CHEN Yuan¹

(1. Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha 410128, China; 2. Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha 410128, China)

Abstract: In order to improve the prediction accuracy of the HIV-1 protease cleavage site, a shear prediction model based on feature selection and support vector machine is proposed. First, by analysis of the cleavage site dataset of 5830 samples, and using absorption minimum redundancy maximum relevance concept, the automatic termination method is employed to select the cleavage site feature vectors. Then, the feature vector is input to a support vector machine for learning and training to build the classification model of splice sites. Finally, simulation is carried out using MATLAB 2004 simulation toolbox. Results show that the proposed model has better prediction accuracy than that of the reference models and literature report. The selected features have good interpretability and biological significance.

收稿日期:2015-12-15.

基金项目:高等学校博士学科点专项科研基金项目(20124320110002);长沙市科技计划项目(K1406018-21).

作者简介:袁哲明(1971-),男,教授,博士生导师.研究方向:大数据分析,生物信息学. E-mail:zhmyuan@sina.com

通信作者:陈渊(1987-),男,讲师.研究方向:大数据分析. E-mail:chenyuan0510@126.com

Key words: biophysics; cleavage site prediction; feature selection; minimal redundancy maximal relevance(mRMR); support vector machine(SVM)

0 引言

获得性免疫缺陷综合症(Acquired immune deficiency syndrome, AIDS)由感染人类免疫缺陷病毒(Human immunodeficiency virus, HIV)引起。HIV 病毒包含两型,在人类中传播的主要是 HIV-1 型。HIV-1 蛋白酶对 HIV-1 病毒的复制必不可少,高效 HIV-1 蛋白酶抑制剂是当前临床治疗艾滋病的重要手段^[1]。现有研究表明^[2],一条蛋白质上 8 个氨基酸残基长度的子序列对 HIV-1 蛋白酶是敏感的。因此,确认可被 HIV-1 蛋白酶剪切的八肽序列对研发 HIV-1 蛋白酶肽类抑制剂意义重大。八肽序列 $P_4 P_3 P_2 P_1 P'_1 P'_2 P'_3 P'_4$ 有 20^8 种组合,剪切点位于 $P_1 P'_1$ 之间,逐一合成验证其可被剪切性费时、费力、费钱,可采用机器学习法。基于给定数据集,如何表征每一条八肽序列是机器学习的首要关键。常用的氨基酸正交编码对八肽序列可产生 160 个 0/1 特征,简单直观,但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,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。但按正交编码,S-N 与 N-W 间差异均为 1。正交编码法在多数情况下结果并不理想^[3]。氨基酸指数数据库(<http://www.genome.jp/aaindex/>)含 20 种天然氨基酸的 531 种理化性质^[4],是氨基酸理化性质的较全面总结,已广泛用于肽序列的表征^[5-7]。

以往的研究多注重于序列表征方法的研究,往往忽略了特征选择的重要性。然而,肽序列经 AA531 表征后存在大量无关与冗余特征。特征选择可降低模型训练耗时,提高预测精度,增强模型解释性。最小冗余最大相关(minimal redundancy maximal relevance, mRMR)是最广泛应用的特征过滤器之一^[8-10]。在因变量为离散变量、自变量(特征)为连续变量时,mRMR 搜索一个最优特征子集 S,使得 S 中自变量与因变量的 t-score(二分类)或 F-score(多分类)均值最大,同时 S 中两两自变量 Pearson 相关系数绝对值|R|-score 均值最小。mRMR 的主要缺点是只能给出特征引入顺序,需通过训练集交叉测试来决定何时引入终止,相当耗时。此外,也存在总体分

布未知时 t-score 与 F-score 应用受限、|R|-score 不能反映非线性相关、相关性测度指标 t-score 或 F-score 与冗余性测度指标|R|-score 不可比等局限。

一些经典分类器,如马尔科夫模型^[11]、隐马尔科夫模型^[12]、K 最近邻分类算法(K-nearest neighbor, KNN)^[13]、人工神经网络^[14]、支持向量机(Support vector machine, SVM)^[13, 15]等都曾用于蛋白质序列分类的研究。SVM 是一种基于结构风险最小的有监督机器学习方法,其结合氨基酸序列信息,近年来被广泛用于蛋白质的亚细胞定位^[16]、功能预测^[17]、相互作用^[18]、功能分类^[19]等蛋白质序列分类研究。与其他分类器相比,SVM 能有效避免过拟合和局部最小等问题,且泛化推广能力更加优异。

本文构建了一个据作者所知迄今最全的 HIV-1 蛋白酶剪切位点数据集,以 AA531 表征肽序列,根据 mRMR 最小冗余最大相关理念,基于 dCor 测度与冗余分摊,提出了一种可自动终止特征引入的特征选择新方法 dCor-share,最后以支持向量分类(Support vector classification, SVC)实施预测,获得了满意的独立预测精度。

1 特征选择和支持向量机的剪切位点预测模型

1.1 特征选择

每条八肽序列以氨基酸指数数据库中所含 531 种理化性质(AA531)直接表征,按序排列,每个样本含 4248 个初始特征^[6]。设 Y 为离散型因变量,连续型自变量(特征)集合 $\Omega = \{X_1, X_2, \dots, X_i, \dots, X_m\}$,集合长度(元素个数) $|\Omega| = m$ 。目前已引入特征集合为 S。记 $\Omega_S = \Omega - S$,则 mRMR 引入下一个特征的标准 F-test correlation difference(FCD)为^[20]:

$$\max_{X_i \in \Omega_S} [F(X_i, Y) - \frac{1}{|S|} \sum_{X_j \in S} |R(X_i, X_j)|] \quad (1)$$

或 F-test correlation quotient(FCQ):

$$\max_{X_i \in \Omega_S} \{F(X_i, Y)/[\frac{1}{|S|} \sum_{X_j \in S} |R(X_i, X_j)|]\} \quad (2)$$

式中: $F(X_i, Y)$ 为单因素不等重复方差分析的 F-score(二分类时 $F = t^2$); $R(X_i, X_j)$ 为 Pearson 相关系数。

一般而言, FCQ 优于 FCD^[20]。本文采用 FCQ-mRMR 作为参比特征选择方法, 设定条件重要性排序后引入特征数上限 $B=100, 200, 300$ 。每引入一个特征, 对训练集以 SVC 进行 5-fold 交叉测试, 直至特征数达到上限 B 。取交叉测试精度最高对应的特征子集为最优特征子集, 用于后续独立测试。

距离相关 dCor 通过计算样本本身的欧几里得距离来衡量变量间的相关程度^[21]。对两个随机变量 X, Y , 其总体距离系数 dCor 定义为:

$$\text{dCor}(X, Y) = \begin{cases} \sqrt{\frac{V^2(X, Y)}{\sqrt{V^2(X)V^2(Y)}}}, & V^2(X)V^2(Y) > 0 \\ 0, & V^2(X)V^2(Y) = 0 \end{cases} \quad (3)$$

$\text{dCor} \in [0, 1]$, 无需预知总体分布, 能探测变量间的非线性或非单调关系, 并有较好的统计势^[21], 可同时替代 F-score 与 $R(X_i, X_j)$ 使之具有可比性。

根据 mRMR 最小冗余最大相关理念, 本文基于 dCor 测度与冗余分摊研究了一种可自动终止特征引入的特征选择新方法 dCor-share。对 S 中的某个已引入特征 X_i , 其冗余分摊后的得分为:

$$\text{dCor-share}(X_i) = \frac{\text{dCor}(X_i, Y)}{\sum_{X_j \in S} \text{dCor}(X_i, X_j)} \quad (4)$$

则 S 中所有特征经冗余分摊后的总得分为:

$$\text{dCor-share}(S) = \sum_{X_i \in S} \frac{\text{dCor}(X_i, Y)}{\sum_{X_j \in S} \text{dCor}(X_i, X_j)} \quad (5)$$

设下一个引入特征为 X_{next} , 记 $D = S + \{X_{\text{next}}\}$, 显然 $|D| = |S| + 1$ 。则 dCor-share 引入下一个最优特征的标准为:

$$\max_{X_{\text{next}} \in \Omega_S} [\text{dCor-share}(D)] = \sum_{X_i \in D} \frac{\text{dCor}(X_i, Y)}{\sum_{X_j \in D} \text{dCor}(X_i, X_j)} \quad (6)$$

终止引入特征标准为:

$$\text{dCor-share}(D) \leq \text{dCor-share}(S) \quad (7)$$

为验证冗余特征分摊策略的有效性, 本文以 F-score、 $|R|$ -score 取代 dCor-share 中的 dCor, 得到另一参比特征选择方法 F-test correlation

share(FC-share)。其引入下一个最优特征的标准为:

$$\max_{X_{\text{next}} \in \Omega_S} [\text{FC-share}(D)] = \sum_{X_i \in D} \frac{F(X_i, Y)}{\sum_{X_j \in D} R(X_i, X_j)} \quad (8)$$

终止引入特征标准为:

$$\text{FC-share}(D) \leq \text{FC-share}(S) \quad (9)$$

dCor-share、FC-share 两种方法由自编 Matlab 程序实现并经验证通过。

1.2 分类器

支持向量机(SVM)通过找到一个最优分类超平面将所有训练样本划分两类, 即

$$\begin{aligned} y_i \{[\Psi(x_i), \omega] + b\} &\geq 1 \\ i = 1, 2, \dots, n \end{aligned} \quad (10)$$

式中: n 为训练样本的数量。

对于一个分类超平面, 参数 (ω, b) 不唯一确定, 因此一定有一对 (ω, b) 保证式(10)成立, 设 $[\Psi(x_i), y_i]$ 与分类超平面最小距离为 $1/\|\omega\|$, 允许存在一些误分类的点, 这样式(10)变为^[22]:

$$\begin{aligned} y_i \{[\Psi(x_i), \omega] + b\} &\geq 1 - \zeta_i \\ i = 1, 2, \dots, n \end{aligned} \quad (11)$$

式中: ζ_i 为负松弛变量, $\zeta_i = 0$ 时, 表示完全线性可分。

对于一个线性不可分问题, 那么就需要将其转为一个优化问题, 再找到最优分类超平面, 通过引入惩罚因子 $C > 0$, 有

$$\begin{cases} \min \Psi(\omega) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n \zeta_i \\ \text{s. t. } \begin{cases} y_i \{[\Psi(x_i), \omega] + b\} \geq 1 - \zeta_i \\ i = 1, 2, \dots, n \end{cases} \end{cases} \quad (12)$$

引入 Lagrange 算子 α_i , 将式(12)转为如下的优化问题:

$$\begin{cases} \max W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \Psi(x_i) \Psi(x_j) \\ \text{s. t. } \begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases} \end{cases} \quad (13)$$

式中: $\omega = \sum_{i=1}^l \alpha_i \Psi(x_i)$ 。

支持向量机的分类判别函数为:

$$f(x) = \omega \Psi(x) + b = \sum_{i \in SV} \alpha_i \Psi(x_i) \Psi(x) + b \quad (14)$$

特征空间的内积(核函数)为:

$$K(x, y) = \sum_i \Psi_i(x) \Psi_i(y) \quad (15)$$

高斯核定义为:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (16)$$

式中: σ 为高斯分布宽度。

支持向量机的分类判别函数为:

$$f(x) = \omega \Psi(x) + b = \sum_{i \in SV} \alpha_i K(x_i, x) + b \quad (17)$$

剪切位点预测是一个多分类问题, 支持向量机为求解两分类问题方法, 为此采用“一对一”方法建立剪切位点预测的分类器。设剪切位点预测的训练样本集为: $[y_1, x_1, \mu(x_1)], \dots, [y_n, x_n, \mu(x_n)]$, $x_i \in \mathbf{R}^n$ 表示样本的特征, $y_i \in \{-1, 1\}$ 表示剪切位点类别, $\mu(x_i)$ 表示样本的特征集合, 则剪切位点预测的最优分类面的最优解为:

$$\begin{cases} \Phi(\omega, \zeta) = \frac{1}{2} \|\omega\|^2 + C \left[\sum_{i=1}^n \mu(x_i) \zeta_i \right] \\ \text{s. t. } \begin{cases} y_i [(\omega^T, z_i) + b] - 1 + \zeta_i \geq 0 \\ \zeta_i \geq 0 \end{cases} \end{cases} \quad (18)$$

相应核函数最优分类面的判别函数式为:

$$\begin{cases} f(x) = \operatorname{sgn} \left[\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b \right] \\ \text{s. t. } \begin{cases} 0 \leq \alpha_i \leq \mu(x_i) C \\ i = 1, 2, \dots, n \end{cases} \end{cases} \quad (19)$$

2 仿真实验

2.1 数据集和模型评价标准

收集的 HIV-1 蛋白酶剪切位点数据集包括 Dataset-746^[23], Dataset-1625^[24], Dataset-3272^[25], Dataset-885^[26] 和 Dataset-947^[27]。去除冗余与矛盾样本, 得到整合数据集 Dataset-5830, 含 990 个正样本, 4840 个负样本。按 3 : 1 随机划分训练集和独立测试集, Train(+)742 个样本, Train(-)3630 个

样本, 248 个样本, Test(+)1210 个样本, 重复 5 次。分类器采用 LibSVM 3.1 软件包, Matlab2014 仿真工具箱^[22]。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。独立测试评价标准采用总正确预测率(ACC)和曲线下面积(AUC)^[28]。

2.2 实验结果

对数据集 Dataset-5830, 以氨基酸正交编码, 每样本得 160 个 0/1 特征, SVC 5 次独立测试 ACC 平均数为 $88.0658\% \pm 0.4172\%$ 。以 AA531 编码, 每样本得 4248 个初始特征, SVC 5 次独立测试 ACC 平均数为 $89.2654\% \pm 0.5754\%$ 。实验结果表明, AA531 编码优于正交编码。

表 1 是 Dataset-5830 数据集上 AA531 编码不同特征选择方法 5 次独立测试的 SVC 预测表现和保留特征数。可见, 随设定条件重要性排序后引入特征数上限 B 增大, FCQ-mRMR 保留特征数大幅度增加, 而预测精度 ACC 和 AUC 上升缓慢; 当 $B=300$ 时, 平均保留特征数达 257.0, 而平均 ACC、AUC 分别为 0.9034、0.9080。改用冗余特征分摊策略, FC-share 平均保留特征数下降为 100.6, 而平均 ACC、AUC 分别上升至 0.9199、0.9144。结果表明, 冗余分摊策略有效。在冗余特征分摊策略基础上进一步改用 dCor 测度, dCor-share 以最少的平均保留特征(68.6/4248)获得了最高的平均 ACC(0.9232)和 AUC(0.9289), 表明改用 dCor 测度能够进一步提升预测精度, 同时, 能显著减少保留特征个数。

由于 Dataset-5830 数据集是本文首次构建, 因此本文选用以往研究较多的 Dataset-1625 和 Dataset-885 数据集。同时, 以往研究采用的特征编码、特征选择方法、分类器很不统一, 故表 2 仅整体比较 dCor-share-SVC 与文献报道结果。可见, 在两个数据集上 dCor-share-SVC 的预测表现均优于文献报道。

表 1 不同特征选择方法的预测表现

Table 1 Prediction performance of different feature selection methods

特征选择方法	特征数	ACC(Mean \pm Std)	AUC(Mean \pm Std)
No selection	4248	0.8927 ± 0.0058	0.9045 ± 0.0132
FCQ-mRMR, B=100	94.2 ± 7.8	0.8908 ± 0.0055	0.8874 ± 0.0251
FCQ-mRMR, B=200	179.2 ± 17.9	0.9026 ± 0.0049	0.9048 ± 0.0130
FCQ-mRMR, B=300	257.0 ± 31.0	0.9034 ± 0.0054	0.9080 ± 0.0112
FC-share	100.6 ± 14.0	0.9199 ± 0.0051	0.9144 ± 0.0034
dCor-share	68.6 ± 11.1	0.9232 ± 0.0043	0.9289 ± 0.0075

表2 dCor-share-SVC与文献报道结果比较
Table 2 Compared dCor-share-SVC with reported results

数据集	特征编码	特征选择方法	分类器	保留特征数/ 初始特征数	k-fold CV	ACC	AUC	文献
Dataset-1625	Discretized ternary z^* -scales	—	Rough Sets	40/40	5	0.930	0.940	[24]
Dataset-1625	OETMAP	—	LSVM	240/240	10	0.950	0.980	[29]
Dataset-1625	AA531	dCor-share	SVC	54/4248	5	0.963	0.983	本文
Dataset-885	AA4	—	Inf. Comf.	32/32	10	0.892	—	[26]
Dataset-885	0/1	C-FS-SVM	SMO	12/160	5	0.919	0.910	[30]
Dataset-885	AA531	dCor-share	SVC	36/4248	5	0.938	0.975	本文

2.3 特征选择方法比较

mRMR 的局限如下:

(1)当前的 mRMR 不适于因变量 Y 为连续型变量的情形。

(2)相关性测度 t-score 或 F-score 在总体呈非正态分布时应用受限。

(3)冗余性测度 $| R(X_i, X_j) |$ 不能反映非线性冗余。

(4)相关性测度与冗余性测度在同一公式(1)或公式(2)中不可比。对相关性测度,若特征对 Y 有用,总有 $F\text{-score} > 1$;而 $| R(X_i, X_j) |$ 恒小于或等于 1。例如, $F(X_A, Y) = 3, \frac{1}{| S |} \sum_{X_j \in S} | R(X_A, X_j) | = 0.9; F(X_B, Y) = 2, \frac{1}{| S |} \sum_{X_j \in S} | R(X_B, X_j) | = 0.1$ 。

按 FCD 则 X_A (得分 2.1) 优先 X_B (得分 1.9) 入选,按 FCQ 则 X_B (得分 20) 优先 X_A (得分 3.33) 入选。可见,FCD 对去冗余作用甚微,FCQ 则过分放大了冗余的危害。

(5)以 FCD 为例,假定当前已引入最优特征集合为 S,则这 $| S |$ 个特征的平均相关性-平均冗余性为:

$$\text{Score}(\bar{S}) = \left[\frac{1}{| S |} \sum_{X_i \in S} F(X_i, Y) - \frac{1}{| S |^2} \sum_{X_i, X_j \in S} | R(X_i, X_j) | \right] \quad (20)$$

现引入下一个最优特征 X_{next} , $D = S + \{X_{\text{next}}\}$,则这 $| D |$ 个特征的平均相关性-平均冗余性为:

$$\text{Score}(\bar{D}) = \left[\frac{1}{| S |} \sum_{X_i \in D} F(X_i, Y) - \frac{1}{| S |^2} \sum_{X_i, X_j \in D} | R(X_i, X_j) | \right] \quad (21)$$

后引入的特征一般相关性要小,采用均值比

较通常有 $\text{Score}(\bar{S})$ 大于 $\text{Score}(\bar{D})$,如图 1 所示。第一个特征与后面的特征其得分不可比,故没有在图 1 中显示。随着特征个数的增加,特征得分会保持下降,直到引入上限为止。因此,mRMR 仅能给出特征的条件重要性排序,不能自动终止特征引入。

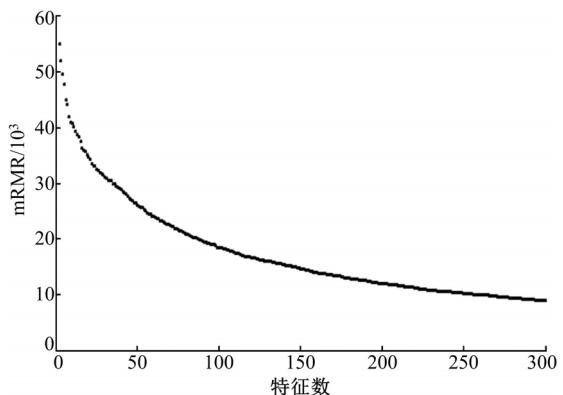


图 1 第一次重复试验 mRMR 特征得分

Fig. 1 mRMR score of features in first replicate

dCor-share 的优缺点如下:

(1)适用于因变量 Y、自变量 X 为连续型变量或二值变量的情形。

(2)相关性测度与冗余性测度统一为 dCor,因而在同一公式中可比。

(3)dCor 不受总体是否呈正态分布所限,且能有效探测变量间的非线性冗余/相关。

(4)假定当前已引入最优特征集合为 S,则这 $| S |$ 个特征经冗余分摊后的总相关性为:

$$\text{dCor-share}(S) = \sum_{X_i \in S} \frac{\text{dCor}(X_i, Y)}{\sum_{X_j \in S} \text{dCor}(X_i, X_j)} \quad (22)$$

现引入下一个最优特征 X_{next} , $D = S + \{X_{\text{next}}\}$,则这 $| D |$ 个特征经冗余分摊后的总相关性为:

$$dCor-share(D) = \sum_{X_i \in D} \frac{dCor(X_i, Y)}{\sum_{X_j \in D} dCor(X_i, X_j)} \quad (23)$$

需要注意的是, D 中每个特征经冗余分摊后的相关性得分在 X_{next} 引入后将被刷新。因而, 随着特征引入, 特征子集经冗余分摊后的总相关性得分存在极大值, 如图 2 所示。因此, $dCor-share$ 无需设定引入特征上限, 无需经交叉测试即可自动终止特征引入, 甚为省时方便。其主要缺陷是不适用于因变量 Y 、自变量 X 为多类(大于等于 3)离散变量的情形。 $dCor-share$ 的有效性尚需更多高维特征选择实例结果的支持。

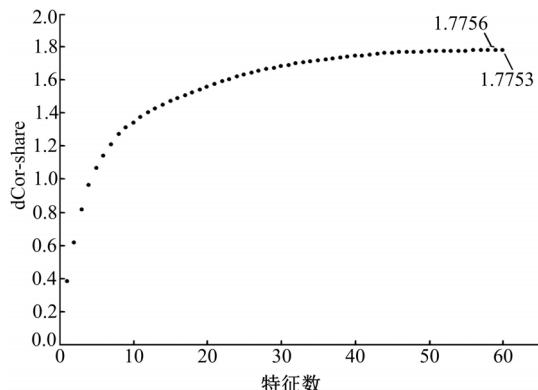


图 2 第一次重复试验 $dCor-share$ 特征子集总得分

Fig. 2 $dCor-share$ score of feature subsets in first replicate

2.4 共有特征分析

好的特征选择方法应具鲁棒性, 即针对同一数据集的多次随机抽样, 所选特征应有较好的重现性。对 Dataset-5830 数据集, 当抽样分数为 0.75(4372/5830)时, 表 1 中 $dCor-share$ 5 次重复试验所选保留特征数依次为 59、70、66、61、87, 共有特征 33 个, 共有率 48.10% (33/68.6), 表明 $dCor-share$ 有较好的重现性。共有特征对解释 HIV-1 蛋白酶可剪切性与各位点残基及其理化性质的关系更具可靠性价值。已有研究表明^[31], 20 种天然氨基酸的 531 种理化性质可归为 6 类。

表 3 结果显示, 33 个共有特征主要集中于 P_1 、 P'_1 、 P'_2 位点, 而 HIV-1 蛋白酶是由两条含有 99 个氨基酸的肽链组成的同型二聚体, 具有 C2 对称轴, 酶的两个亚单位的一端包含一个移动性较高的所谓“挡板结构”^[32], 它由一个反平行片层和一个螺旋延展至底物的结合位点 P_1 和 P'_2 形成, 文献^[23,24]报道这两个位点是影响 HIV-1 型蛋白酶可剪切性的重要位点, 这与本文的发现大致相同; 同时, 除了 P_1 、 P'_1 、 P'_2 位点之外的其余位点也不容忽视, 符合 HIV-1 蛋白酶可剪切序列无固定模式或保守序列的现有认知, 也体现了本研究的难度。

表 3 共有保留特征的频次分布

Table 3 Frequency distribution of common retain features

性质分组	P_4	P_3	P_2	P_1	P'_1	P'_2	P'_3	P'_4	合计
A	1	1	2	1	0	3	1	0	9
B	0	0	0	0	1	0	1	0	2
C	1	0	0	1	0	1	0	1	4
H	1	2	1	2	4	2	2	1	15
O	0	0	0	0	0	1	0	0	1
P	0	0	0	1	0	0	0	1	2
合计	3	3	3	5	5	7	4	3	33

注:A 为 alpha and turn propensities; B 为 beta propensity; C 为 Composition; H 为 Hydrophobicity; P 为 Physicochemical properties; O 为 other properties。

HIV-1 蛋白酶可剪切性主要与各位点残基的疏水性、Alpha 螺旋与转角属性有关。其中, 在 P_1 和 P'_1 位点上主要为疏水性, 目前, 绝大多数文献^[23,24,32]记载的 HIV-1 蛋白酶抑制剂都在 P_1 和 P'_1 位点具有疏水性, 在 HIV-1 型蛋白酶中与 P_1 和 P'_1 位点结合的氨基酸残基^[33] 主要包括 Arg8, Leu23, Asp25, Gly27, Gly48, Gly49, Ile50, Thr80, Pro81, Val82, 这些氨基酸残基也都具有疏水性特征, 这与本文发现在 P_1 和 P'_1 位点上主要为疏水性相吻合。能够在 HIV-1 型蛋白酶中与 P'_2 位点结合的氨基酸残基主要包括 Ala28, Asp29, Asp30, Val32, Ile47, Leu76, Ile84, 这些氨基酸残基通常比对应结合位点的残基要小, 且在 HIV-1 型蛋白酶内部通过折叠形成一个“袋”状结构^[33], 据此可知, 本文发现在 P'_2 位点主要与 Alpha 螺旋与转角属性有关这一推论具有较高的可信度。

3 结束语

为了提高 HIV-1 型蛋白酶剪切位点的预测准确性, 提出一种基于特征选择和支持向量机的剪切位点预测模型。实验结果表明, 本文模型在特征最少的条件下, HIV-1 型蛋白酶剪切位点预测精度优于参比模型。同时, 所用特征选择方法

能有效地选择出具有较好生物学意义的特征,提升了模型的可解释性。与其他特征选择方法相比,该特征选择方法最大的优势在于,考虑特征子集的综合得分,而不是单个特征的综合评价得分,在特征选择的过程中存在极值,可以自动终止特征引入,因此该特征选择方法在高维具有广泛应用前景。

参考文献:

- [1] Rodríguez-Barrios F, Gago F. HIV protease inhibition: limited recent progress and advances in understanding current pitfalls[J]. Current Topics in Medicinal Chemistry, 2004, 4(9): 991-1007.
- [2] Schechter I, Berger A. On the size of the active site in proteases. I. papain[J]. Biochemical and Biophysical Research Communications, 1967, 27(2): 157-161.
- [3] Nanni L, Lumini A. A new encoding technique for peptide classification[J]. Expert Systems with Applications, 2011, 38(4): 3185-3191.
- [4] Kawashima S, Pokarowski P, Pokarowski M, et al. AAindex: amino acid index database, progress report 2008[J]. Nucleic Acids Research, 2008, 36(Sup. 1): 202-205.
- [5] Cao D S, Xu Q S, Liang Y Z. Progy: a tool to generate various modes of Chou's PseAAC[J]. Bioinformatics, 2013, 29(7): 960-962.
- [6] 韩娜,袁哲明,陈渊,等. 基于高维特征非线性筛选的HLA-A^{*}0201限制性CTL表位预测[J]. 物理化学学报, 2013, 29(9): 1945-1953.
Han Na, Yuan Zhe-ming, Chen Yuan, et al. Prediction of HLA-A^{*} 0201 restricted Cytotoxic T Lymphocyte epitopes based on high-dimensional descriptor nonlinear screening [J]. Acta Phys-Chim Sin, 2013, 29(9): 1945-1953.
- [7] 李咏,周玮,代志军,等. 基于序列特征筛选与支持向量回归预测蛋白质折叠速率[J]. 物理化学学报, 2014, 30(6): 1091-1098.
Li Yong, Zhou Wei, Dai Zhi-jun, et al. Predicting the protein folding rate base on sequence feature screeing and support vector regression [J]. Acta Phys-Chim Sin, 2014, 30(6): 1091-1098.
- [8] Li B Q, Huang T, Liu L, et al. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network [J]. PLoS one, 2012, 7(4): e33393.
- [9] Li Y, Wang M, Wang H, et al. Accurate in species-specific acetylation sites by integrating protein sequence-derived and functional features[J]. Scientific Reports, 2014, 4: 5765.
- [10] Ma X, Sun X. Sequence-based predictor of ATP-binding residues using random forest and mRMR-IFS feature selection[J]. Journal of Theoretical Biology, 2014, 360: 59-66.
- [11] Chou K C. Prediction of human immunodeficiency virus protease cleavage sites in proteins[J]. Analytical Biochemistry, 1996, 233(1): 1-14.
- [12] Jayavardhana R G L, Palaniswami M. Cleavage knowledge extraction in HIV-1 protease using hidden Markov model[C]// Intelligent Sensing and Information Processing, Chennai, India, 2005: 469-473.
- [13] Nanni L, Lumini A. Mpps: an ensemble of support vector machine based on multiple physicochemical properties of amino acids [J]. Neurocomputing, 2006, 69(13): 1688-1690.
- [14] Niu B, Lu L, Liu L, et al. HIV-1 protease cleavage site prediction based on amino acid property[J]. Journal of Computational Chemistry, 2009, 30(1): 33-39.
- [15] Sarda D, Chua G H, Li K B, et al. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties[J]. BMC Bioinformatics, 2005, 6(1): 1-12.
- [16] Chou K C, Cai Y D. Using functional domain composition and support vector machines for prediction of protein subcellular location[J]. Journal of Biological Chemistry, 2002, 277(48): 45765-45769.
- [17] Cai Y D, Liu X J, Xu X B, et al. Support vector machines for predicting protein structural class[J]. BMC Bioinformatics, 2001, 2(1): 1-5.
- [18] Bock J R, Gough D A. Predicting protein-protein interactions from primary structure[J]. Bioinformatics, 2001, 17(5): 455-460.
- [19] Cai C, Han L Y, Ji Z L, et al. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence [J]. Nucleic Acids Research, 2003, 31(13): 3692-3697.
- [20] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205.
- [21] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distance

- [J]. *The Annals of Statistics*, 2007, 35(6): 2769-2794.
- [22] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2007, 2(3):389-396.
- [23] You L, Garwica D, Rögnvaldsson T. Comprehensive bioinformatics analysis of the specificity of human immunodeficiency virus type 1 protease [J]. *Journal of Virology*, 2005, 79(19): 12477-12486.
- [24] Kontjevskis A, Wikberg J E, Komorowski J. Computational proteomics analysis of HIV-1 protease interactome[J]. *Proteins: Structure, Function, and Bioinformatics*, 2007, 68(1): 305-312.
- [25] Rögnvaldsson T, Etchells T A, You L, et al. How to find simple and accurate rules for viral protease cleavage specificities [J]. *BMC Bioinformatics*, 2009, 10(1):1-17.
- [26] Jaeger S, Chen S S. Information fusion for biological prediction[J]. *Journal of Data Science*, 2010, 8(2): 269-288.
- [27] Impens F, Timmerman E, Staes A, et al. A catalogue of putative HIV-1 protease host cell substrates [J]. *Biological Chemistry*, 2012, 393(9): 915-931.
- [28] Fawcett T. ROC graphs: notes and practical considerations for researchers [J]. *Machine Learning*, 2004, 31: 1-38.
- [29] Gök M, Özcerit A T. A new feature encoding scheme for HIV-1 protease cleavage site prediction [J]. *Neural Computing and Applications*, 2013, 22(7):1757-1761.
- [30] Öztürk O, Aksac A, Elsheikh A, et al. A consistency-based feature selection method allied with linear SVMs for HIV-1 protease cleavage site prediction [J]. *PLoS One*, 2013, 8(8): e63145.
- [31] Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins[J]. *Protein Engineering*, 1996, 9(1): 27-36.
- [32] Poorman B R A, Tomasselli A G, Heinrikson R L, et al. A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base[J]. *Journal of Biological Chemistry*, 2010, 266(22):14554-14561.
- [33] Ezziane Z. Application of artificial intelligence in bioinformatics: a review[J]. *Expert Systems with Applications*, 2006, 30(1): 2-10.