中国科学 B 辑: 化学
2007 年 第 37 卷 第 6 期: 592~596
http://www.scichina.com

《中国科学》杂志社
SCIENCE IN CHINA PRESS

支持向量机和线性判别分析用于禽流感病毒编码蛋白识别

梁桂兆[①②③] 陈泽聪[①②] 杨善彬[②③] 梅 虎[②③] 周 原[②③] 舒 茂[①②]
杨 力[①③] 杨胜喜[①②] 郑小林[②③] 陈国华[①②] 周 鹏[①②] 田菲菲[①②]
廖春阳[①②] 吴世容[①②] 李根容[①②] 李德静[①②] 何 留[①②] 甘孟瑜[①②]
高剑坤[①②③] 陈国平[①③] 王贵学[②③] 龙 莎[①②] 景举华[①②] 曾 晖[①②]
张巧霞[①②] 张梦军[①②③] 杨 娟[①②③] 仝建波[①②] 王娇娜[①②] 刘永红[①②]
李 波[①] 仇亮加[①] 蔡绍皙[①③] 赵 娜[①②] 杨 艳[①②] 苏霞利[①②] 宋 健[②]
陈美霞[②] 陈 刚[①④] 张雪姣[①②] 孙家英[①②] 李经纬[①②] 邓 婕[①②]
彭传友[①②] 李 志[①②] 许罗南[①②] 廖立敏[①②] 吴玉乾[①②] 朱万平[①②]
苏勤亮[①] 卢大军[①] 李 军[①] 黄振虎[①②] 周 萍[①②] 李志良[①②*]

(① 重庆大学化学化工学院, 生物医学工程重庆市重点实验室, 重庆 400044; ② 湖南大学化学生物传感与计量学国家重点实验室, 长沙 410012; ③ 重庆大学生物工程学院, 生物力学与组织工程教育部重点实验室, 重庆 400030; ④ 新加坡国立技术学院生命科学技术中心, 新加坡 139651)

摘要 以支持向量机(SVM)和线性判别分析(LDA)对 200 条禽流感病毒、100 条 B 型流感和 100 条 C 型流感病毒蛋白共 400 条为训练集样本, 从表征序列的 200 个整体与局部变量中以逐步(stepwise)方法选取 24 个变量作为 LDA 模型的输入建立线性识别模型, 病毒蛋白总识别率达 99.8%, 留一法交互检验总识别率为 99.4%. 从原始 200 变量中经主成分分析得 16 个主成分作为 SVM 的输入, 以径向基核函数(RBF)SVM 建立非线性识别模型, 病毒蛋白总识别率为 99.8%, 留一法交互检验总识别率为 99.2%. 以 100 条禽流感、50 条 B 型流感和 50 条 C 型流感病毒编码蛋白质共 200 条为测试集样本, 得 LDA 模型, 对其总识别正确率为 95.4%, SVM 模型对其总识别正确率为 96.5%. 识别结果表明, 两个模型都可较好识别禽流感病毒蛋白, 并且 SVM 对禽流感病毒蛋白的识别结果优于 LDA.

关键词 禽流感病毒 编码蛋白 支持向量机(SVM) 线性判别分析(LDA) 径向基函数(RBF)

流感病毒属正黏液病毒科, 根据核蛋白和基质蛋白抗原性不同, 可分为A、B和C(或称为甲、乙和丙)三个血清型. B型变异缓慢, 只感染人, 流行比较局限. C型很少变异, 在人类和猪中存在, 但病毒相对温和,

收稿日期: 2005-11-27; 接受日期: 2006-03-13
重庆大学研究生创新团队项目科技创新基金(批准号: 200711C1A0010260)、国家"春晖计划"教育部启动基金(批准号: 99-4-4+37)、霍英东基金(批准号: 98-7-6)、重庆直辖市应用基础基金(批准号: 01-3-6)、生物力学与组织修复工程创新引智基地(教技[2005]7 号 111 计划)、湖南大学化学生物传感与计量学国家重点实验室项目(批准号: 20050012)及重大自主创新科技攻关项目(批准号: 03-5-6+04-10-10)资助
* 联系人, E-mail: zlli-cqu@163.com; zlli2662@163.com

不常发生. 其中A型易发生变异, 可引起禽、鸟类、人、猪、马和海洋哺乳动物暴发流行. 禽流感病毒[1,2](AIV)属于A型, 根据血凝素(HA)和神经氨酸酶(NA)亚型不同可分为不同亚型. 到目前已发现 15 种HA抗原(H1~H15)和 9 种NA抗原(N1~N9). 每个病毒含一个HA和NA蛋白. 自 1878 年首次分离至今, AIV一直是严重危害世界各国养禽业发展的头号大敌, 对高致病性(HP)AIV可使鸡和火鸡 100%致死, LP的AIV可引起较小局部感染或不易导致疾病. 目前为止, HP的AIV仅发生在H5 和H7 亚型, 虽然并不是所有H5 和H7 呈现HP.

因AIV抗原及其致病力的易变性, 这便要求采取快速检测, 并用先进分子生物学技术进行病毒鉴定、检疫及免疫保护等措施, 从而保证人和动物安全. 对A、B和C型流感病毒可通过琼脂扩散试验、补体结合试验测出它们之间抗原性差别. 然而用这些方法识别流感病毒较耗时和耗物力. 本室曾对SARS-CoV编码蛋白建立快速识别模型并取得较好结果[3], 我们进一步选择组成AIV编码蛋白质的整体及局部特性诸如序列氨基酸组成、物化和空间结构性质参数, 对AIV和已知完整基因组序列的其他B和C型流感病毒共 400 个编码蛋白样本, 采用LDA和SVM方法建立识别模型, 取得较好结果. 继用 200 条流感病毒序列做外部验证集, 在此基础上试图寻找AIV编码蛋白区别于B和C病毒编码蛋白的总体特征, 希望能为AIV蛋白识别及相关基因组和蛋白质组研究提供一些有益信息和参考依据.

1　原理与方法

1.1　样本来源及序列参数化

从美国GenBank库随机抽取 200 条AIV编码蛋白、100 条B型流感病毒蛋白和 100 条C型流感病毒蛋白共 400 条序列(附表 1, 读者可向作者索取)作为研究对象, 所选流感病毒蛋白序列所含残基从25至774个不等. 研究表明[2,3], HA裂解为HA1 和HA2 是AIV致病的重要因素, 在病毒入侵细胞及决定病毒致病力方面具关键作用. 同时, 构成HA碱性氨基酸的多少和宿主体内蛋白裂解酶的分布是决定病毒致病能力及其在机体内扩散能力的主要因素, 故可认为AIV编码蛋白的氨基酸序列构成、位置及其物化性质和空间结构性质对其特性有显著影响. 鉴于此, 为全面表征序列结构, 从其整体和局部结构特征出发, 收集各种流感病毒蛋白序列的各种参数, 包括: 分子质量, 等电点, 体积, 比容, 立体性质, 电性, 疏水性, 蛋白质的空间二级结构, 氨基酸出现次数与频率及自由能等(附表 2, 读者可向作者索取).

1.2　LDA 建模

LDA[4]根据如下判别函数判别观测值所属类别: $Y=a_0+a_1X_1+a_2X_2+\cdots+a_nX_n$, 其中, Y是判别分数, X_1, X_2, X_3, …, X_n指描述子变量(观测值), a_1, a_2, a_3,…, a_n指各变量系数, 即判别系数. 通过建立一个超平面使空间形成两个区域, 然后判别化合物属于哪一个(区域)类别. LDA由SPSS13.0 完成, 首先令“1”表示AIV蛋白, “2”表示非AIV蛋白(流感B和C). 因描述子变量较多, 故以stepwise方法挑选进入模型的变量, 变量挑选原则以偏F检验对应F值为依据, 当F值大于 3.84 时, 则该变量留在模型中, 当该变量对应F值小于 2.71 时, 则剔除该变量.

1.3　SVM 建模

SVM由Vapnik等发展, 因较强泛化能力引起研究者广泛关注与浓厚兴趣. 其基于结构风险最小化理论寻找最优界面实现最小泛化误差, 而非实现最小化经验误差, 能较好解决小样本、非线性、高维数和局部最小等问题, 且能从原理上解决其他传统机器算法(如神经网络)过拟合现象. 已被成功用于人脸识别, 语音识别, 手写文字识别, 垃圾邮件识别, 金融风险评估, 蛋白质空间结构识别及基因识别等诸多模式识别领域中[5~12]. SVM用于解决线性可分两类样本分类问题的核心思想是找到一个最优分类超平面, 使两类样本的分类间隔最大化. 对于非线性问题, 首先经非线性映射Φ, 将样本映射到一高维特征空间, 然后用线性方法解决. 为此SVM引入核函数: $K(x, x_i) = \Phi(x)\cdot\Phi(x_i)$实现高维映射. 核函数的引入, 可有效避免维数灾难及计算复杂性等问题. 目前常用核函数主要有线性核: $K(x, x_i) = x\cdot x_i$; 多项式核: $K(x, x_i) = (\alpha_1 x\cdot x_i+\alpha_2)^p$; RBF核: $K(x, x_i) = \exp(-\gamma\|x-x_i\|^2)$; Sigmoid核: $K(x, x_i) = \tanh(\alpha_1 x\cdot x_i+\alpha_2)$. 我们选择非线性的RBF核SVM用于建模. SVM计算程序由Matlab7.0软件编写完成.

1.4 模型评价及验证

以分类正确率表示识别结果: $R_{cla}=\{(N_t-N_e)/N_t\}\times100\%$, 式中$N_t$为样本总数; N_e是被错误分类样本数. 同时, 以留一法交互检验验证模型内部预测能力, 但对于一个良好模型外部验证必不可少 [13], 故从美国GenBank挑选与训练集不同的 100 条AIV编码蛋白、50 条B型流感病毒蛋白和 50 条C型流感病毒蛋白序列共 200 条作外部测试集(附表 3, 读者可向作者索取), 验证所建模型的外部识别能力.

2 结果与讨论

2.1 LDA 模型识别结果

经 stepwise 方法, 最终有 24 个变量进入 LDA 模型, 表 1 表示 LDA 模型中变量名称及方程的非标准化系数及标准化系数. 识别结果(附表 1)表明, 对 400 条训练集流感病毒蛋白的总识别率为 99.8%, 留一法交互检验总识别率为 99.4%(表 2). 据方程标准化系数可分析 24 个变量对 AIV 蛋白识别结果影响显著性. 由表 3 看出, W 出现频率, 在三氟乙酸中的保留系数, R 占总质量百分数, Q 占总质量百分数, N 出现次数及θ值(i)等六个变量对应标准化系数绝对值均>5, 故推测以上性质对 AIV 编码蛋白质与 B 和 C 型流感病毒蛋白性质存在较显著影响. 其中标准化系数最大(+10.157)的是 N 出现次数, N 的侧链中的酰胺是一个不解离的极性基团, 但其可与水形成氢键, 故推断随着 N 出现次数的增加, 有形成 AIV 蛋白的倾向. 变量三氟乙酸中的保留系数对应的标准化系数为 8.289, 它是蛋白质的一个疏水特性, 对 AIV 蛋白形成亦具积极影响. 在对 AIV 蛋白形成不利的因素中, 影响最大的是 Q 占总质量百分数, 虽然 Q 侧链中的酰胺可与水形成氢键, 但却受到序列中其他氨基酸残基所占质量分数影响, 其他残基占总质量百分数越大, 对 AIV 形成越有利. 对于变量 R 占总质量百分数, 虽然带正电荷胍基的存在可能使R化学活性较为活跃, 但其所占质量百分数越大对 AIV 形成越不利. 变量θ值(i)对应标准化系数为–5.997, 表明其亦不利于 AIV 蛋白形成. W 出现频率对 AIV 蛋白形成也具不利影响, W 侧链中含有吲哚稠杂环, 故 W 出现频率越高, 可能因其空间阻隔效应而不利于 AIV 蛋白形成. 其他变量影响在此不作分析.

表 1 用于 LDA 建模的 24 个变量及其标准化系数与非标准化系数及剔除变量的 F 值

序号	变量名称	偏 F 检验值	非标准化系数	标准化系数
常数	/	/	1.067	/
1	W 出现频率	130.209	–0.265	–5.292
2	N 原子与 N 原子之间的几何距离总和	12.377	–0.064	–3.214
3	三氟乙酸中的保留系数	269.913	334.507	8.289
4	碱性氨基酸出现次数	83.583	0.276	3.334
5	H 出现次数	11.304	0.091	0.742
6	V 出现频率	6.760	–0.150	–0.862
7	Q 出现次数	98.835	0.425	2.724
8	G 出现次数	58.200	105.144	2.193
9	R 占总质量百分数	305.617	–0.442	–6.223
10	侧链角θ	255.701	1.165	4.032
11	S 出现频率	130.073	0.332	4.023
12	I 出现次数	72.358	0.431	2.774
13	V 出现次数	36.268	–61.539	–1.183
14	Q 占总质量百分数	23.253	–338.692	–7.633
15	N 出现次数	40.011	595.566	10.157
16	C 出现次数	129.147	0.675	2.461
17	平均可及表面积	54.213	–0.223	–2.258
18	M 出现次数	147.924	108.587	1.553
19	M 出现频率	80.158	4.960	2.447
20	S 占总质量百分数	62.693	1.422	2.477
21	D 出现次数	55.766	–1.180	–3.938
22	电荷转移给体能力参数	3.842	1.833	1.747
23	极性氨基酸出现次数	5.354	13.250	0.388
24	θ值(i)	138.692	–21.63	–5.997

表 2　流感病毒蛋白识别结果

模型	项目	结果	+[a)]	–[b)]	总计
SVM_{mm}	数目	+	200	0	200
		–	1	199	200
	百分比	+	100.0	0.0	100.0
		–	0.5	99.5	100.0
LDA_{mm}	数目	+	199	1	200
		–	0	200	200
	百分比	+	99.5	0.5	100.0
		–	0.0	100.0	100.0
SVM_{cv}	数目	+	198	2	200
		–	1	199	200
	百分比	+	99.0	1.0	100.0
		–	0.5	99.5	100.0
LDA_{cv}	数目	+	199	1	200
		–	1	199	200
	百分比	+	99.5	0.5	100.0
		–	0.5	99.5	100.0
SVM_{ext}	数目	+	98	2	100
		–	5	95	100
	百分比	+	98.0	2.0	100.0
		–	5.0	95.0	100.0
LDA_{ext}	数目	+	91	9	100
		–	0	100	100
	百分比	+	91	9.0	100.0
		–	0.0	100.0	100.0

a) “+”代表禽流感病毒蛋白样本, b) “–”代表 B 型与 C 型流感病毒蛋白样本

为验证模型外部的识别能力, 根据识别函数非标准化系数方程, 可得出禽流感组中心得分点为 3.728, 非禽流感组中心得分点为–3.728, 故以“0”为分解点将某序列的 24 个变量带入非标准化系数方程, 若判别分大于 0, 则为 AIV, 若小于 0, 则为非 AIV, 对 200 条非 AIV 蛋白(B 和 C)(识别结果见附表 3)识别率为 95.4%(表 2), 结果显示, 模型具较高外部识别能力.

2.2　SVM 模型识别结果

参数设置对 SVM 回归成败至关重要. 比如, 对于 RBF 核 SVM, 惩罚系数 C, 参数γ, 核函数类型 K 及相应参数选择等. 其中 C 是调控参数, 用于调节最大间隔和最小化训练误差间平衡, 若 C 太小, 训练将导致欠拟合, 若过大, 将导致过拟合, 为使学习过程稳定, 应将 C 设置相对大一些. 参数γ控制 Gaussian 函数振幅即 SVM 泛化能力. 我们根据所建模型对外部预测集所得正确率, 应用单因素轮换法, 确定 SVM 各参数.

研究发现, 以 200 个原始变量作为SVM的输入, 对 400 条训练集和 200 条测试集样本的识别正确率均较低. 故对 200 个变量采取主成分分析 [14]降维, 以各主成分对应的特征值(大于 1)为选取标准. 最终选取 16 个变量(解释了原始变量矩阵 98.2%的方差)作为SVM的输入. 固定γ=0.031, 然后分别取C=5, 10, 30, 50, 100, 300, 600 和 1000, 识别正确率随C变化见图 1, 当C=100 时, 相对最大识别正确率为 96.5%. 固定 C = 100 时, 取γ= 0.125, 0.056, 0.031, 0.020, 0.014, 0.01, 0.008 和 0.001, 图 2 识别正确率随γ变化表明, 当γ= 0.031 时, 相对最大识别正确率为 96.5%. 故将 RBF核SVM回归参数设置为: C = 100, 核函数为$K(x, x_i)=\exp(-0.031\|x-x_i\|^2)$, 对 400 条训练集样本总识别

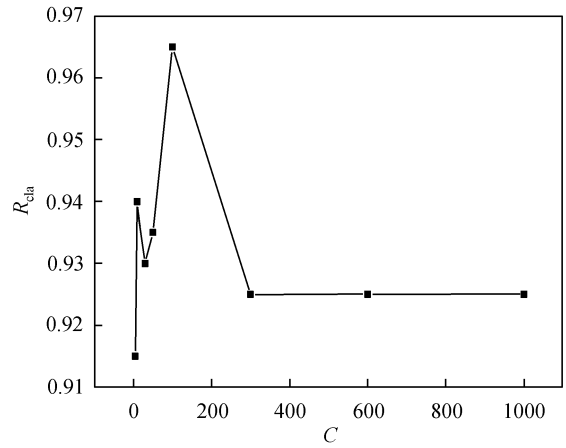

图 1　识别正确率随 C 变化图(γ= 0.031)

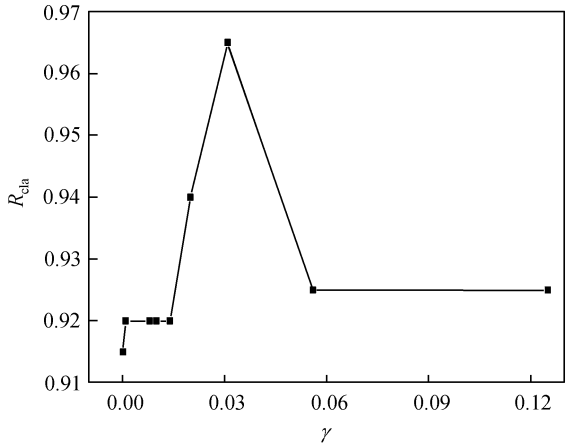

图 2 识别正确率随γ变化图(C = 100)

正确率为 99.8%(附表 1), 与 LDA 方法相当, 虽然交互检验预测正确率(99.2%)稍低于 LDA(99.4%), 但对 200 条外部验证集蛋白识别率(96.5%)(识别结果见附表 3)却高于 LDA(95.4%)(统计结果见表 2), 整体识别结果表明, SVM 优于 LDA.

3 结论

快速识别AIV对于预防AIV及采取有效措施控制AIV蔓延具重要实践意义 [15~19]. 从序列的整体和局部特征出发收集了序列200种性质参数, 分别用LDA和非线性的RBF核SVM建立AIV识别模型, 结果表明, 两个模型都可较好识别AIV蛋白与非AIV蛋白. 从识别结果比较可知, SVM对AIV蛋白识别结果稍优于LDA, 但SVM的参数选择相对复杂, 在参数选择方法有必要作进一步探讨. AIV识别模型的建立可为研究者了解AIV和非AIV蛋白特征, 进一步为相关基因组与蛋白质组研究提供有益参考依据.

参 考 文 献

1 甘孟候. 禽流感(第二版). 北京: 中国农业出版社, 2002. 15—79
2 Garten W, Klenk H D. Understanding influenza virus pathogenicity. Trends in Microbiology, 1999, 7(3): 99—100[DOI]
3 梅虎, 孙立力, 周原, 熊清, 李志良. SARS-CoV 编码蛋白质的识别. 科学通报, 2005, 49(15): 1503—1506)
4 Kachigan S K. Statistical Analysis. New York: Radius Press, 1986. 264—285
5 Burbidge R, Trotter M, Buxton B, et al. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. Comp Chem, 2001, 26: 5—14[DOI]
6 Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. J Mol Biol, 2001, 308: 397—407[DOI]
7 Hong S J, Weiss S M. Advances in predictive models for data mining. Pattern Recogn Lett, 2001, 22: 55—61[DOI]
8 Belousov A I, Verzakov S A, Frese J von. A flexible classification approach with optimal generalization performance: support vector machines. Chemom Intell Lab Syst, 2002, 64: 15—25[DOI]
9 Gao J B, Gunn S R, Harris C J. SVM regression through variational methods and its sequential implementation. Neurocomputing, 2003, 55: 151—167[DOI]
10 Cortes C, Vapnik V. Support vector networks. Mach Learn, 1995, 20: 273—293
11 Flach P A. On the state of the art in machine learning: A personal review. Artificial Intelligence, 2001, 131: 199—222[DOI]
12 Sánchez A V D. Advanced support vector machines and kernel methods. Neurocomputing, 2003, 55: 5—20[DOI]
13 Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and inerpretation of QSPR models. QSAR Comb Sci, 2003, 22: 69—77[DOI]
14 Kim D, Lee I-B. Process monitoring based on probabilistic PCA. Chemon Intell Lab Syst, 2003, 67: 109—123[DOI]
15 Guan Y, Poon L L M, Cheung C Y. H5N1 influenza: A protean pandemic threat. Proc Natl Acad Sci, 2004, 101(21): 8156—8161 [DOI]
16 Chen H, Deng G, Li Z, Tian G, Li Y, Jiao P, Zhang L, Liu Z, Webster R G, Yu K. The evolution of H5N1 influenza viruses in ducks in southern China. Proc Natl Acad Sci USA, 2004, 101: 10452—10457[DOI]
17 Guan Y, Shortridge K F, Krauss S, et al. Molecular characterization of H9N2 influenza viruses: Were they the donors of the "internal" genes of H5N1 viruses in Hong Kong? Proc Natl Acad Sci USA, 1999, 96: 9363—9367[DOI]
18 Chen H, Smith G J D, Li K S. Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. Proc Natl Acad Sci USA, 2006 103: 2845—2850[DOI]
19 Lin Y P, Shaw M, Gregory V. Avian-to-human transmission of H9N2 subtype influenza A viruses: Relationship between H9N2 and H5N1 human isolates. Proc Natl Acad Sci USA, 2000, 97: 9654—9658[DOI]