研究简报

基于SVM-RFE的水稻抗病基因筛选

付 媛[1], 王 岩[1,2], 周 柚[1], 张 帆[1], 王珏鑫[1], 梁艳春[1],
(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林大学 数学学院, 长春 130012)

摘要: 提出一种改进的回归特征消去支持向量机特征选择方法(SVM-RFE)对水稻的抗病基因进行筛选. 实验结果表明: 在预测得到的20个与水稻抗病/敏感相关基因中, 有3个基因与已知的水稻抗病基因紧密相关; 2个基因与已知的水稻抗病基因有一定的相关性. 通过该方法能找到影响水稻生长状态(正常/染病)的基因.

关键词: 回归特征消去支持向量机; 基因筛选; 水稻抗病
中图分类号: TP39 **文献标志码**: A **文章编号**: 1671-5489(2011)06-1101-04

Disease Resistance Related Gene Screening in *Oryza sativa* Using SVM-RFE

FU Yuan[1], WANG Yan[1,2], ZHOU You[1], ZHANG Fan[1], WANG Yu-xin[1], LIANG Yan-chun[1]
(1. *College of Computer Science and Technology*, *Jilin University*, *Changchun* 130012, *China*;
2. *College of Mathematics*, *Jilin University*, *Changchun* 130012, *China*)

Abstract: An improved support vector machine recursive feature extraction (SVM-RFE) algorithm was used to screen the disease resistance genes. In the 20 important genes, we found that 3 of them have strong relation to the disease resistance as reported and 2 of them have relation to the stress response. It shows that this method can find out which genes could impact the rice growth status (normal/disease). It might provide a guide on finding other unknown rice disease resistance/sensibility genes in biology.

Key words: support vector machine recursive feature elimination (SVM-RFE); gene screening; rice disease resistance

由于水稻抗病基因所用的基因表达数据具有小样本、高维度的特点, 因此使用传统的机器学习方法分析存在一定的困难. 而支持向量机(SVM)在解决小样本、高维度数据集时表现出许多特有的优势, 并被广泛应用于基因表达数据分析中. Furey等[1]结合遗传算法和支持向量机提取最优特征, 准确率达99%; 基于网络的SVM算法, 对临床相关致病基因的分类, 也有很高的准确率[2]. 基于SVM的递减前向选择(iterative reduced forward selection, IRFS)方法, 在选择特征子集前进行特征选择, 该方法消除了某些非相关特征, 从而减少了计算量, 加快了特征选择过程.

1 实 验

1.1 数据来源 基因表达数据源于NCBI的GEO[3-4]公共数据库. 目前, 水稻基因表达数据样本较少,

收稿日期: 2011-02-26.

作者简介: 付 媛(1987—), 女, 汉族, 硕士研究生, 从事生物信息学的研究, E-mail: fu.yuan730@gmail.com. 通讯作者: 王 岩(1978—), 男, 汉族, 博士, 副教授, 从事生物信息学的研究, E-mail: wy6868@hotmail.com; 梁艳春(1953—), 男, 汉族, 博士, 教授, 博士生导师, 从事生物信息学的研究, E-mail: ycliang@jlu.edu.cn.

基金项目: 国家自然科学基金(批准号: 61073075; 60903097)、国家高技术研究发展计划863项目基金(批准号: 2009AA02Z307)、教育部博士点基金(批准号: 20090061120094)和吉林省青年基金(批准号: 20090116).

水稻疾病数据仅有24个GSE系列,且一半GSE系列少于10个样本;水稻染病和抗病的实验数据仅有3个GSE系列:GSE6124,GSE6125和GSE16142. 而在这3个GSE中,前2组仅各有3个和6个样本,样本过少,不适于本文研究. 因此,本文选择GSE16142作为抗病数据研究目标,该系列包含两个子系列:GSE16140和GSE16141,其中GSE16140由9个名为TW16的样本组成,属于水稻东格鲁病[5]的抗病品种;GSE16141包含12个名为台中在来一号(*Oryza sativa* cv. Taichung Native 1)的样本,属于水稻东格鲁病的易感品种. 数据由日本农业生物资源研究所提供,并发布于GEO. 所用的水稻基因组芯片均来自美国安捷伦科技(Agilent Technologies)公司,芯片类型编号为GPL7252,包含45 911个探针. 本文将这些探针所对应的基因ID号作为特征,进行SVM-RFE特征提取分析. 实验得到由基因ID号组成的特征集后,在GO数据库中查找对应的蛋白描述,明确其生物学意义.

1.2 *数据预处理* 由于基因表达数据具有高维度、高噪声、高冗余等特点,因此,必须进行数据预处理,否则实验结果的精度会受到影响. 在移除数据质量较差(含0较多)的噪声列后,用t-test方法过滤掉不相关基因,并用归一化方法将剩余数据归一到$[-1,1]$区间.

1.2.1 *t-test方法* 为了判定不同基因表达数据间的差异是否具有统计显著性,本文采用t-test检验方法[6]进行假设检验,检验两类样本的某一特征是否具有相同的均值,其零假设为均值相等,即两组值间没有显著差异. 若选择显著性水平α,则置信区间为$1-\alpha$. 若对两类样本中某一特征T,有$p>\alpha$,则拒绝零假设,表明其均值没有显著变化. 本文选择$\alpha=0.001$得到845个特征,这些特征对应的表达数据在两类样本中有明显差异.

1.2.2 *数据归一化* 数据的归一化即将所有数据通过某种函数的作用变换到同数值区间内. 通过数据的归一化可将数据转换为无量纲的纯量从而统一,使样本有了相同的统计分布特性,这样可较方便地进行相关系数计算,分析数据中的统计学意义[7].

本文采用如下形式的线性转换函数:

$$x_{ij}=\frac{2(x_{ij}-\min\limits_{i}x_{ij})}{\max\limits_{i}x_{ij}-\min\limits_{i}x_{ij}}-1,\qquad j=1,2,\cdots,N, \tag{1}$$

该线性函数将基因表达数据映射到$[-1,1]$内.

1.3 *改进的SVM-RFE方法*

1.3.1 *选择核函数* 支持向量机的核心是引入核函数,以避免直接进行内积计算. 因此,在高维空间中不需得到映射后高维函数的具体形式,只需知道核函数即可求解,由于核函数映射是非线性的,从而得到了非线性支持向量机的算法. 本文实验选择最简单的线性核函数.

1.3.2 *改进的SVM-RFE方法* 基于SVM,Guyon等[8]首次提出SVM-RFE(support vector machine recursive feature elimination)方法,该方法利用回归特征消去(RFE)法,逐个消去基因,取得了较好效果,成为基因选择中的经典算法.

传统的SVM-RFE方法首先训练SVM分类器;然后按评估分数从大到小对特征进行排序,移去列表最后的特征,经迭代,重新计算特征排序列表,再移去列表的最后特征,直到预测准确率达到最高,或已经达到需要的特征数. 此时,所得到的最小特征子集即为最优特征子集.

本文根据数据的特点,对21个样本(9个抗病样本,12个易感样本)使用留一交叉验证法(LOOCV)进行21次回归特征消去. 每次使用20个数据进行实验,留下1个数据不参与实验,21个数据均有1次不参与实验,这样可以获得更好的泛化能力和更高的精度. 本文设定最优特征子集的最大容量为30,选出每次实验中特征排序列表的前30个,根据各特征出现的次数和排名位置进行加权求和并再次排序,构成最终的特征排序列表. 权值定义为

$$w_i=\begin{cases}(M+1-i)/M, & 1\leqslant i\leqslant M,\\ 0, & i>M,\end{cases} \tag{2}$$

其中M为选取特征的个数,本文取$M=30$. 若某个特征f在第k个特征列表中的位置为c_k,则权值为$W_f=\sum\limits_{k}w_{ck}(1\leqslant k\leqslant T)$,其中$T$为实验次数. 算法流程如图1所示.

2　实验结果

通过以上改进的 SVM-RFE 过程，本文得到前 20 个特征探针 ID 如下（以 ID 号为标记，按重要程度排序）：{15338，14364，33936，4512，8950，2247，29830，27662，12490，42899，32612，45081，5804，34415，12781，13687，27324，1089，34635，9121}.

特征探针 ID 对应的基因对该数据集的分类预测有较大贡献[9]，这些探针所对应的基因在两类样本中表达差异较大，可能与水稻抗东格鲁球形病毒密切相关. 为了评估这些基因是否具有抗病功能的生物学意义，本文检索了 GO 等数据库，找到以下3 个已知与水稻抗病紧密相关的基因.

1）ID 为 8950 的探针，对应 bZIP 转录因子 HBP-1a. bZIP 转录因子是普遍存在于动植物及微生物中的一类转录因子，以识别核心序列 ACGT 的顺式作用元件，如 CACGTG（G 盒）、GACGTC（C 盒）、TACGTA（A 盒）等，其中 G 盒元件普遍存在于受 ABA、生长素、茉莉酸和水杨酸诱导的基因中，表明它与植物的抗逆性有关，参与植物防卫反应[10].

2）ID 为 29830 的探针，转录名为 *OsWRKY*80 的蛋白质. *OsWRKY*80是 WRKY 转录因子超家族的一员，该转录因子是一类与植物抗病防卫反应相关的转录因子，为植物中特有的超基因家族，该家族成员都含有高度保守的 WRKYGQK 结构域和 CzHH/C 锌指结构基序，并通过与顺式作用元件 W. box：(T)(T)TGAC(C/T)相结合识别目标基因[11]. WRKY 基因在植物的防卫反应中具有重要的调控作用，会在病毒、细菌或其他非生物胁迫中被诱导，具有较高的表达. 在干涉转基因植株中内源 *OsWRKY*80 基因的诱导表达被抑制，转基因植株的抗病性与 *OsWRKY*80 基因的表达呈一定的正相关性，表明 *OsWRKY*80基因可能作为正调控因子参与水稻防卫反应[12].

3）ID 为 27324 的探针，转录蛋白质谷胱甘肽 *s*-转移酶（GSTU6）. GSTU6 是一组具有多种生理功能的蛋白质，在机体有毒化合物的代谢、保护细胞免受急性毒性化学物质攻击中具有重要作用. 谷胱甘肽 *s*-转移酶可催化亲核性的谷胱甘肽与各种亲电子外源化合物的结合反应. 许多外源化合物在生物转化第一相反应中极易形成某些生物活性中间产物，它们可与细胞生物大分子重要成分发生共价结合，对机体产生损害. 谷胱甘肽与其结合后，可防止发生此种共价结合，具有解毒作用. 且 GSTU6 能在转录水平对植物进行调节，参与抵御过量氧的复杂反应，与植物的抗病性密切相关[13].

以上 3 种基因与水稻的抗病性存在较强的相关性，其余部分基因的表达虽然没有明显与抗病性相关，但在植物的应激反应中具有一定的作用，如 SLR1 基因.

ID 为 2247 的探针，对应 SLR1 基因，能转录 *s* 位点相关糖蛋白（SLR1）. SLR1 蛋白属于 DELLA 蛋白家族，研究表明，DELLA 蛋白是 GA 信号转导途径中的重要调控因子，GA 通过降解 DELLA 蛋白以消除抑制效果，实现自身的激素功能[14]. GA 是一种广泛存在的植物激素，具有促进植物生长的作用，种子萌发、下胚轴和茎的伸长、叶片的扩展、花和种子的发育等过程都有 GA 的参与. ID 为 27324 的探针，对应 NF-YB3（核转录因子 YB 亚基-3）基因，当内质网产生应激反应时，NF-YB3 得到表达，形成一个复杂的转录过程，上调内质网应激诱导的基因表达[15]. ID 为 33936 的探针，对应 EARLY flowering 4 基因，是一种与花诱导相关并能调节与生物中相关基因表达的基因.

这些重要基因（每个探针对应唯一一个基因），在 21 次 SVM-RFE 过程中的出现次数如图 2 所示

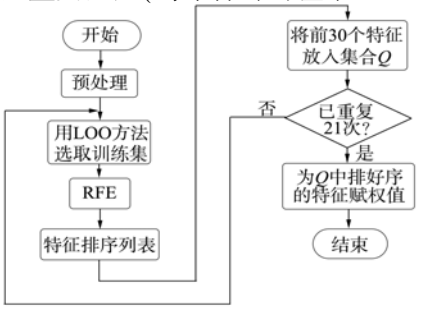

图 1　改进的回归特征消去支持向量机算法流程

Fig. 1　Flowchart of improved SVM-RFE algorithm

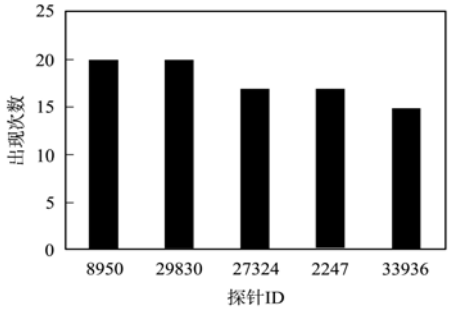

图 2　基因出现次数

Fig. 2　Gene appearances during experiments

(图2中用探针号表示基因). 由图2可见, 在传统方法的实验中, 某些不重要的基因也会排在特征序列表的前面, 实验结果稳定性较差; 而改进后的方法, 平衡了基因在特征序列表中出现的位置和次数, 突出了重要的基因, 提高了实验结果的稳定性.

由以上实验可见, SVM-RFE在一定程度上能对水稻基因表达数据进行分类. 由于水稻属于禾本科模式植物, 因此本文方法也适用于小麦等其他禾本科作物, 也可应用于对其他植物的基因分类上. 本文研究水稻的抗病性, 由于相关样本数量过少, 在一定程度上影响了实验精度和效果.

参 考 文 献

[1] Furey T S, Cristianini N, Duffy N, et al. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data [J]. Bioinformatics, 2000, 16(10): 906-914.

[2] TANG Li-juan, JIANG Jian-hui, WU Hai-long, et al. Variable Selection Using Probability Density Function Similarity for Support Vector Machine Classification of High-Dimensional Microarray Data [J]. Talanta, 2009, 79(2): 260-267.

[3] MAO Yong, ZHOU Xiao-bo, PI Dao-ying, et al. Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree with Gene Selection [J]. J Biomed Biotechnol, 2005(2): 160-171.

[4] ZHOU Xin, Wu X Y, MAO Ke-zhi, et al. Fast Gene Selection for Microarray Data Using SVM-Based Evaluation Criterion [C]//2008 IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia: IEEE Computer Society, 2008: 386-389.

[5] LIU Hua, MA Wen-li, ZHENG Wen-ling. GEO (Gene Expression Omnibus): High-Throughput Gene Expression Database [J]. Chinese Journal of Biochemistry and Molecular Biology, 2007, 23(3): 236-244. (刘华, 马文丽, 郑文岭. GEO (Gene Expression Omnibus): 高通量基因表达数据库 [J]. 中国生物化学与分子生物学报, 2007, 23(3): 236-244.)

[6] YU Hai-lang, MA Wen-li, ZHENG Wen-ling. Data Mining Procedures Using GEO (Gene Expression Omnibus) [J]. China Biotechnology, 2007, 27(8): 96-103. (余海浪, 马文丽, 郑文岭. 用于基因数据挖掘的基因表达数据库GEO [J]. 中国生物工程杂志, 2007, 27(8): 96-103.)

[7] Smyth G K, Speed T. Normalization of cDNA Microarray Data [J]. Methods, 2003, 31(4): 265-273.

[8] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification Using Support Vector Machines [J]. Machine Learning, 2002, 46(1/2/3): 389-422.

[9] ZHOU Jin-bo. Prediction of Disease-Resistant Gene in Rice Based on SVM-RFE [D]: [Master's Degree Thesis]. Changchun: College of Computer Science and Technology, Jilin University, 2010. (周金博. 基于SVM-RFE的水稻抗病基因预测 [D]: [硕士学位论文]. 长春: 吉林大学计算机科学与技术学院, 2010.)

[10] LUO Sai-nan, YANG Guo-shun, SHI Xue-hui, et al. On the Application of Transcription Factor to Plant Stress Resistance [J]. Journal of Hunan Agricultural University: Natural Sciences, 2005, 31(2): 219-223. (罗赛男, 杨国顺, 石雪晖, 等. 转录因子在植物抗逆性上的应用研究 [J]. 湖南农业大学学报: 自然科学版, 2005, 31(2): 219-223.)

[11] Baldi P, Long A D. A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized *t*-Test and Statistical Inferences of Gene Changes [J]. Bioinformatics, 2001, 17(6): 509-519.

[12] LI Nan-yi, CAI Rong-yao, GUO Ze-jian. The Disease Resistance of Rice Regulated by *OsWRKY*80 Gene [J]. Acta Agriculturae Shanghai, 2009, 25(3): 14-18. (李南羿, 柴荣耀, 郭泽建. *OsWRKY*80基因参与调控水稻抗病反应研究 [J]. 上海农业学报, 2009, 25(3): 14-18.)

[13] REN Yan-jiao, WANG De-ping, WANG Yan, et al. Prediction of Disease-Resistant Gene in Rice Based on SVM-RFE [C]//Proceedings of 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI). Yantai: The IEEE Engineering in Medicine and Biology Society, 2010: 2343-2346.

[14] Hussain A, PENG Jin-rong. DELLA Proteins and GA Signaling in Arabidopsis [J]. Plant Growth Regul, 2003, 22(2): 134-140.

[15] LIU Jian-xiang, Howell S H. bZIP28 and NF-Y Transcription Factors Are Activated by ER Stress and Assemble into a Transcriptional Complex to Regulate Stress Response Genes in *Arabidopsis* [J]. The Plant Cell, 2010, 22(3): 782-796.