doi: 103969/j. issn. 0490-6756. 2016. 03. 032

基于 SVM 的革兰氏阴性菌分泌系统蛋白识别方法

余乐正[1,2], 赵柳青[1], 陈　曼[1], 罗杰斯[2], 柳凤娟[1]
(1. 贵州师范学院化学与生命科学学院,贵阳 550018;
2. 四川大学化学学院,成都 610065)

摘　要: 本文提出了一种基于 SVM 快速识别革兰氏阴性菌分泌系统蛋白的方法. 该方法以氨基酸组成和位置特异性得分矩阵为最优特征集,充分考虑了蛋白质的序列信息及进化信息. 实验结果表明,本文提出的方法对革兰氏阴性菌分泌系统蛋白具有较好的预测性能,可作为细菌分泌系统研究的有益补充.

关键词: 革兰氏阴性细菌;分泌系统蛋白;SVM;位置特异性得分矩阵
中图分类号: Q811.4　　**文献标识码**: A　　**文章编号**: 0490-6756(2016)02-0443-05

A SVM based approach to identification of Gram-negative bacterial secretion system proteins

YU Le-Zheng[1,2], *ZHAO Liu-Qing*[1], *CHEN Man*[1], *LUO Jie-Si*[2], *LIU Feng-Juan*[1]
(1. School of Chemistry and Life Science, Guizhou Normal College, Guiyang 550018, China;
2. College of Chemistry, Sichuan University, Chengdu 610065, China)

Abstract: A SVM based approach is proposed to rapidly identify Gram-negative bacterial secretion system proteins. With the optimization feature set consisted of amino acid composition (AAC) and position specific scoring matrix (PSSM), this method adequately takes sequence and evolution information of proteins into account. Experiments show that this method has a good performance on prediction of Gram-negative bacterial secretion system proteins, which served as a useful complement to the study of bacterial secretion system.

Key words: Gram-negative bacteria; secretion system proteins; SVM; position specific scoring matrix

1　引言

蛋白质分泌在维持细胞正常的生理活动中发挥着重要作用,可发生在所有生物体中. 所有在细胞内合成,再被分泌到其它细胞器、细胞外环境及其它细胞内起作用的蛋白质,统称为分泌蛋白. 目前,在革兰氏阴性菌细胞中已发现了至少八种分泌系统,根据外膜分泌机制它们分别被命名为第一类分泌系统(Type Ⅰ secretion system, T1SS)至第八类分泌系统(Type Ⅷ secretion system, T8SS)[1]. 这些分泌系统大都由一些具有特殊功能的蛋白质、多肽组成. 对革兰氏阴性菌分泌系统的深入研究表明,这些分泌系统介导蛋白质分子转运穿过双层膜,不仅关系到细菌的生存,更与细菌的致病性密切相关[2]. 通过各种设计精巧的分泌系统,细菌可将毒性因子和效应子释放到细胞外环境中或直接注入宿主体内,从而与宿主进行分子交流[3]. 因此,对细菌分泌系统蛋白进行深入研究,不仅有助

收稿日期:2014-12-06
基金项目:贵州省科学技术基金(黔科合 J 字[2014]2134 号);贵州师范学院校级博士项目(12BS024);贵州师范学院校级大学生科研项目(2013DXS061)
作者简介:余乐正(1984－),男,四川井研人,博士,副教授,研究方向为生物信息学、化学计量学. E-mail: xinyan_scu@126.com

于全面理解、认识、分析和解释蛋白质的分泌机制及各种生理和病理现象,也为疾病的诊断与治疗、新药的研发提供了更多参考.

目前,分泌系统和分泌蛋白已成为生物学领域研究的热点和难点.对于大部分的分泌系统,其分泌机制已基本解释清楚了,且已开发出一些能识别分泌系统蛋白的方法.该类方法通常只能识别某一特定类型的分泌系统蛋白(如第三类分泌系统蛋白)[4-7],或并非专门针对分泌系统蛋白[8],此外大规模识别方法还很少被报道.针对这一问题,本文基于 SVM 算法以及严格的特征筛选程序,构建了一个分类预测器以快速识别革兰氏阴性菌分泌系统蛋白.对于训练集,本方法对革兰氏阴性菌分泌系统蛋白的预测准确率为 87.15%.对于检测本方法实际预测性能的测试集,其预测准确率为 83.66%.此外,另一个公共独立测试集被用于进一步评估本方法的实际预测性能.

2 材料与方法

2.1 材料

本文所用的实验数据均来自于 Pundhir 和 Kumar 的研究[9].其中,正样本集由 1977 条细菌分泌系统蛋白组成,而负样本集则包含 1932 条未定位于细胞壁上的非分泌系统蛋白.此外,用于独立测试的数据集共包含 112 条分泌系统蛋白和 88 条非分泌系统蛋白.

2.2 分类方法

1995 年,Vapnik 等人在统计学习理论的基础上提出了一种新的机器学习算法——支持向量机(SVM)[10].支持向量机主要具有以下优点:由于采用了结构风险最小化原则,能较好地解决小样本学习问题,具有很好的泛化能力[11,12];通过定义核函数,从而将非线性问题转化为线性问题来解决,降低了算法的复杂度,适合于处理非线性问题;支持向量机算法是一个凸优化问题,这就保证其局部最优解一定是全局最优解.正是由于这些优点,支持向量机已被广泛应用于处理生物学领域的各种分类问题,在细菌分泌系统蛋白的识别中也有成功的应用[9].

2.3 模型的性能评估参数

本研究为两类分类问题:一类为细菌分泌系统蛋白(目标类);一类为非细菌分泌系统蛋白.模型的识别效果通过以下 4 个参数进行描述:灵敏度(Sensitivity, SE),特异性(Specificity, SP),准确率(Accuracy, ACC)和 Matthew 相关系数(Matthew′s correlation coefficient, MCC)[13].

$$SE = \frac{TP}{TP + FN} \tag{1}$$

$$SP = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{4}$$

上式中,TP 为真阳性,指属于目标类的正确识别样本数;FP 表示假阳性,指属于目标类的错误识别样本数;TN 表示真阴性,指不属于目标类的正确识别样本数;FN 表示假阴性,指不属于目标类的错误识别样本数.

3 实验部分

3.1 训练集与测试集

由于原始数据中部分蛋白质序列的相似度或同源性很高,冗余的序列信息可能导致模型的过拟合现象,从而降低模型的稳定性和泛化能力.为提高实验数据质量,本文中所有蛋白质序列均通过 BLASTCLUST 软件[14]进行多序列比对,以确保数据集中任意两条蛋白质序列间的相似度均≤25%.去冗余后,正样本集中 1165 条细菌分泌系统蛋白被保留,而负样本集中 1021 条非细菌分泌系统蛋白被保留.随后,在正负样本集中随机抽取 70%的样本作为训练集,而剩余的 30%则作为测试集[9,15].最终,训练集共包含 816 条分泌系统蛋白和 715 条非分泌系统蛋白,测试集则包含 349 条分泌系统蛋白和 306 条非分泌系统蛋白.

3.2 特征提取与表征

蛋白质序列特征的提取是基于计算的蛋白质分类研究中最基本的问题,也是决定模型实际预测性能的关键因素.不同类型的蛋白质,其具有的序列、结构等特征也大不相同,因此特征的正确选取对于模型的预测性能具有决定性作用.本文中,我们分别采用氨基酸组成、氨基酸理化性质及位置特异性得分矩阵以表征蛋白质中氨基酸的序列信息、理化特性及进化信息.

3.2.1 氨基酸组成

氨基酸组成(Amino acid composition, AAC)反映了 20 种天然氨基酸在蛋白质中出现的频率信

息，每条蛋白质序列被转化为一个 20 维的数字向量.

3.2.2　理化性质　本文采用亲水性、等电点、极性、溶剂可及表面积、转移能量和侧链氨基酸体积这 6 个氨基酸理化性质，以表征蛋白质序列中氨基酸残基的结构与功能特征，电荷性质及溶解度等. 考虑到序列中氨基酸残基间的邻接效应，我们通过自协方差变换(Auto covariance，AC)[16]将不等长的蛋白质序列转换为等长的向量.

作为一种常用的统计工具，自互协方差常用于数值向量的分析，并已广泛用于蛋白质的分类研究. 自互协方差会产生两种变量，即相同描述符间的自协方差变量和不同描述符间的互协方差变量. 已有研究结果证实，自协方差变量是自互协方差变量中最重要的组分，其维数也远小于自互协方差变量的维数[17]. 因此，本文也只采用自协方差变量来描述蛋白质序列中氨基酸残基间的邻接效应，并通过公式(5)、(6)进行计算.

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}(P_{i,j} - \frac{1}{n}\sum_{i=1}^{n}P_{i,j}) \times (P_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n}P_{i,j}) \tag{5}$$

$$D = \lg \times q \tag{6}$$

上述两个公式中，参数 lag 表示蛋白质 P 中两个被考虑的氨基酸间的距离，lg 是最大的 lag 值，j 是第 j 个描述符值，i 是蛋白质序列中的第 i 个氨基酸，n 是蛋白质 P 的序列长度，q 是描述符数，而 D 是自协方差变量数. 由于研究中共用到了 6 个理化性质，且 lg 值经实验后取 5，故通过自协方差变换后，每条蛋白质序列被转换为一个 30 维的向量.

3.2.3　位置特异性得分矩阵　在蛋白质的分类研究中，进化信息通常对模型的预测性能具有较大贡献[18]. 作为一种对进化信息进行描述的常用特征，位置特异性得分矩阵(Position specific scoring matrix，PSSM)常被用于表征蛋白质样本. 本文采用 PSI－BLAST 程序(期望值的阈值为 10^{-3})对 Swiss－Prot 数据库进行搜索，经过 3 次迭代后，得到了每条蛋白质的位置特异性得分矩阵. 每个位置特异性得分矩阵含有 $L\times 20$ 个元素，如公式(7)所示：

$$P_{\mathrm{PSSM}} = \begin{bmatrix} R_{1\to1} & R_{1\to2} & \cdots & R_{1\to j} & \cdots & R_{1\to20} \\ R_{2\to1} & R_{2\to2} & \cdots & R_{2\to j} & \cdots & R_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{i\to1} & R_{i\to2} & \cdots & R_{i\to j} & \cdots & R_{i\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{L\to1} & R_{L\to2} & \cdots & R_{L\to j} & \cdots & R_{L\to20} \end{bmatrix} \tag{7}$$

上述公式中 L 表示查询蛋白质序列 P 的长度，(1,2，・・・20)表示按字母顺序排列的 20 种常见氨基酸. $R_{i\to j}$ 表示蛋白质序列 P 中第 i 位的氨基酸在进化过程中突变为第 j 类氨基酸的频率，如公式(8)所示：

$$R_{i\to j} = \frac{R^0_{i\to j} - \frac{1}{20}\sum_{j=1}^{20}R^0_{i\to j}}{\max(R^0_{i\to j}) - \min(R^0_{i\to j})}(i = 1,2,\cdots,L; j = 1,2,\cdots,20) \tag{8}$$

其中，$R^0_{i\to j}$ 表示 PSI－BLAST 程序的原始得分值.

最后，通过公式(9)将所有蛋白质的 PSSM 矩阵转换为等长的向量：

$$\overline{P}_{PSSM} = [\overline{R}_1\ \overline{R}_2 \cdots \overline{R}_j \cdots \overline{R}_{20}]^T (j = 1,2,\cdots,20) \tag{9}$$

其中 T 是转置运算符，$\overline{R_j}$ 表示 PSSM 矩阵中第 j 列的均值. 经过这些处理后，每条蛋白质序列被转换为一个 20 维的向量.

3.3　蛋白质替代模型

基于以上特征，本文共构建了 6 个蛋白质替代模型，以作为支持向量机的数据输入. 其中，模型 1 仅由氨基酸组成(AAC)构成，模型 2 仅由自协方差变量(AC)构成，模型 3 仅由位置特异性得分矩阵(PSSM)构成. 模型 4 为伪氨基酸组成(Pseudo-amino acid composition，PseAAC)替代模型，由氨基酸组成和自协方差变量通过以下公式融合而成：

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+6\times\lg} \end{bmatrix} \tag{10}$$

其中，

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20}f_i + \omega\sum_{j=1}^{6\lg}\theta_j}, (1 \leqslant u \leqslant 20) \\ \dfrac{\omega\theta_{(u-20)}}{\sum_{i=1}^{20}f_i + \omega\sum_{j=1}^{6\lg}\theta_j}, (21 \leqslant u \leqslant 20+6\lg) \end{cases} \tag{11}$$

模型 5 为伪位置特异性得分矩阵(Pseudo-position specific scoring matrix, PsePSSM)替代模型,由位置特异性得分矩阵和自协方差变量通过公式(10)、(11)融合而成. 模型 6 也是一个融合替代模型,由氨基酸组成和位置特异性得分矩阵构成.

3.4 模型的构建

作为一种功能十分强大的机器学习算法,支持向量机(SVM)已被广泛应用于生物学的各个领域. 本文中,我们采用 libsvm 2.89 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)工具箱来构建 SVM 模型. 模型的核函数为径向基函数,并通过网格搜索法对其正则化参数 C 和核函数宽度参数 γ 进行优化. 在统计预测中,独立数据集测试、样本分组测试(5 倍交叉验证法和 10 倍交叉验证法)和留一法(Jackknife test)常被用于检测各种计算方法的预测性能,而留一法被认为是最客观的[19]. 因此,本文也采用留一法构建预测模型.

4 结果与分析

4.1 特征的选择及替代模型的确定

基于 3.3 节提到的 6 个替代模型,本文分别构建了 6 个 SVM 模型,训练结果如表 1 所示.

表 1 不同替代模型对训练结果的影响

Tab. 1 Performance of different substitution models

模型	C	γ	准确率
模型 1	2.0	0.5	81.25
模型 2	8.0	2.0	65.32
模型 3	2.0	2.0	88.44
模型 4	2.0	0.5	83.47
模型 5	8.0	0.5	86.02
模型 6	8.0	0.5	88.50

由表 1 可知,模型 6 的训练结果最好,模型 3 的训练结果略低于模型 6,模型 5、模型 4 和模型 1 的准确率均超过了 80%,而模型 2 的准确率仅为 65.32%,表明自协方差变量提供的蛋白质信息量最少. 模型 4 的准确率优于模型 1,表明随着模型所含蛋白质信息量的增加,模型的预测性能有所提升. 模型 5 的训练结果略低于模型 3,则进一步表明自协方差变量的加入,一方面增加了蛋白质信息量,但另一方面也导致信息冗余,且冗余信息使得模型 5 的训练结果有所降低. 模型 6 也产生了冗余信息,这使得模型 6 的训练结果仅略优于模型 3 的. 考虑到模型 6 所包含的蛋白质信息量更多,且训练结果最好,本文选择模型 6 而不是模型 3 作为最终的蛋白质替代模型,并依此构建了最终的 SVM 模型.

4.2 模型的实际应用

通过 3.1 节构建的测试集,我们对 SVM 模型的实际预测性能进行了检测. 349 条细菌分泌系统蛋白质中 306 条被本方法正确地识别,而 306 条非细菌分泌系统蛋白中共有 242 条被准确预测. 因此,对于该测试集,本方法的灵敏度为 87.68%,特异性为 79.08%,准确率为 83.66%,Matthew 相关系数为 67.19%.

为了进一步评估本方法对细菌分泌系统蛋白的实际预测性能,本文通过 2.1 节提到的独立测试集再次对 SVM 模型进行了检测. 对于该独立测试集,共有 103 条细菌分泌系统蛋白和 79 条非细菌分泌系统蛋白被本方法正确识别,其灵敏度为 91.96%,精密度为 89.77%,准确率为 91.00%,Matthew 相关系数为 81.74%,预测结果令人满意.

5 结 语

本文深入分析了革兰氏阴性菌分泌系统蛋白的各种特征,基于支持向量机构建了一个预测模型以快速准确识别革兰氏阴性菌分泌系统蛋白. 实验结果表明,本方法对革兰氏阴性菌分泌系统蛋白具有较好的预测性能,可作为一种有益的补充工具应用于蛋白质分泌系统的研究. 对于细菌分泌系统蛋白的研究,下一步工作的重点是如何构建多元分类预测器,以实现对不同类型的分泌系统蛋白的快速准确识别.

参考文献:

[1] Desvaux M, Hébraud M, Talon R, *et al*. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue[J]. Trends Microbiol, 2009, 17(4): 139.

[2] 张丽勃,许景升,徐进,等. Ⅵ型分泌系统与革兰氏阴性病原细菌致病性的关系[J]. 植物保护,2011, 37(3): 7.

[3] 冯洁,徐进,许景升,等. Ⅱ型分泌系统与植物病原细菌致病性的关系[J]. 农业生物技术学报,2007, 15(3): 365.

[4] Löwer M, Schneider G. Prediction of type Ⅲ secretion signals in genomes of gram-negative bacteria [J]. PLoS One, 2009, 4(7): e5917.

[5] Arnold R, Brandmaier S, Kleine F, *et al*. Sequence-based prediction of type Ⅲ secreted proteins [J]. PLoS Pathog, 2009, 5(4): e1000376.

[6] Yang Y, Zhao J Y, Morgan R L, *et al*. Computational prediction of type Ⅲ secreted proteins from gram-negative bacteria [J]. BMC Bioinformatics, 2010, 11(Suppl 1): S47.

[7] Yang X J, Guo Y Z, Luo J S, *et al*. Effective identification of gram-negative bacterial type Ⅲ secreted effectors using position-specific residue conservation profiles[J]. PLoS One, 2013, 8(12): e84439.

[8] Pundhir S, Vijayvargiya H, Kumar A. PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes[J]. In Silico Biol, 2008, 8(3-4): 223.

[9] Pundhir S, Kumar A. SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems[J]. Bioinformation, 2011, 6(10): 380.

[10] Vapnik V. Statistical Learning Theory[M]. New York: Wiley, 1998.

[11] 陈思羽，宁芊，周新志，等. DAG-SVM 的结构优化研究及其在故障诊断中的应用[J]. 四川大学学报：自然科学版，2015,52(2)：299.

[12] 徐亮亮，傅德胜. 基于模糊支持向量机的夏季雨型的预报方法研究[J]. 四川大学学报：自然科学版，2013，50(6)：1230.

[13] Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme [J]. Biochim Biophys Acta, 1975, 405(2): 442.

[14] Altschul S F, Madden T L, Schäffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Res, 1997, 25(17): 3389.

[15] Luo J S, Yu L Z, Guo Y Z, *et al*. Functional classification of secreted proteins by position specific scoring matrix and auto covariance[J]. Chemometr Intell Lab, 2012, 110(1): 163.

[16] Wold S, Jonsson J, Sjöström M, *et al*. DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures [J]. Anal Chim Acta, 1993, 277(2): 239.

[17] Guo Y Z, Yu L Z, Wen Z N, *et al*. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences[J]. Nucleic Acids Res, 2008, 36(9): 3025.

[18] Shen H B, Chou K C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM[J]. Protein Eng Des Sel, 2007, 20(11): 561.

[19] Chou K C, Zhang C T. Prediction of protein structural classes[J]. Crit Rev Biochem Mol Biol, 1995, 30(4): 275.