doi: 10.3969/j.issn.0490-6756.2020.04.025

一种快速准确区分Ⅲ型、Ⅳ型分泌效应蛋白的计算方法

柳凤娟[1]，杨　庆[2]，陈　倩[2]，余乐正[2,3]，李益洲[3]
(1. 贵州师范学院地理与资源学院，贵阳 550018；2. 贵州师范学院化学与材料学院，贵阳 550018；
3. 四川大学化学学院，成都 610065)

摘　要：通过Ⅲ型、Ⅳ型、Ⅵ型分泌系统，革兰氏阴性菌可将效应蛋白直接注入宿主体内，并导致宿主感染各种疾病. 由于Ⅲ型、Ⅳ型分泌效应蛋白均属非经典分泌蛋白，且它们可能具有相似的序列模体或进化保守性，故两者之间难于区分. 基于支持向量机和伪位置特异性得分矩阵，本文提出了一种可快速准确识别革兰氏阴性菌Ⅲ型、Ⅳ型效应蛋白的计算方法. 测试集实验结果表明，本方法对Ⅲ型、Ⅳ型效应蛋白具有较好的分类效果，可作为辅助工具用于分泌效应蛋白的进一步研究.

关键词：革兰氏阴性菌；分泌效应蛋白；支持向量机；位置特异性得分矩阵；留一法
中图分类号：O604　　**文献标识码**：A　　**文章编号**：0490-6756(2020)04-0781-05

A fast and accurate computational approach for the distinction of type Ⅲ and Ⅳ secreted effector proteins

LIU Feng-Juan[1], *YANG Qing*[2], *CHEN Qian*[2], *YU Le-Zheng*[2,3], *LI Yi-Zhou*[3]
(1. School of Geography and Resources, Guizhou Education University, Guiyang 550018, China;
2. School of Chemistry and Materials Science, Guizhou Education University, Guiyang 550018, China;
3. College of Chemistry, Sichuan University, Chengdu 610065, China)

Abstract: Through type Ⅲ, Ⅳ, Ⅵ secretion systems, Gram-negative bacterial effector proteins can be directly injected into host cells, which causes the hosts infected with various diseases. Because both of type Ⅲ, Ⅳ secreted effector proteins belong to the non-classically secreted proteins (NCSPs), and may have similar sequence motifs or evolutionary conservation profiles, it is hard to distinguish them from each other. Based on support vector machine (SVM) and pseudo position specific scoring matrix (PsePSSM), a computational approach is proposed to fastly and accurately classify the type Ⅲ and Ⅳ effector proteins of Gram-negative bacteria. The test results show that this approach has a good effect on the classification of type Ⅲ and Ⅳ effector proteins, and could be used as a supplementary tool for further studies of secreted effector proteins.

Keywords: Gram-negative bacteria; Secreted effector protein; Support vector machine; Position specific scoring matrix; Leave-one-out

收稿日期：2019-04-22
基金项目：国家科技部和国家自然科学基金奖励补助资金(黔科合平台人才[2017]5790-07)；贵州省普通本科高等学校青年科技人才成长项目(黔教合 KY 字[2016]219)；贵州师范学院校级博士项目(2018BS002)
作者简介：柳凤娟(1984－)，女，汉族，博士，副教授，研究方向为环境分析与生物信息学. E-mail: morose1984@163.com
通讯作者：余乐正. E-mail: xinyan_scu@126.com

1 引 言

蛋白质分泌在协调细菌与其周围环境间相互作用中发挥着重要作用. 通过各种分泌系统,细菌可将自身合成的蛋白质释放到细胞外,或直接注入真核宿主及相邻细菌细胞内,进而发挥其毒力效应[1]. 目前,经实验证实的革兰氏阴性菌分泌系统至少已有 9 种,它们分别被称为Ⅰ型至Ⅸ型分泌系统[2]. 在这些分泌系统中,Ⅰ型、Ⅱ型、Ⅴ型分泌系统可将各种酶转运到周围环境中,而Ⅲ型、Ⅳ型、Ⅵ型分泌系统则可将各种效应蛋白直接运输到宿主细胞内,其对应的分泌蛋白也分别被命名为Ⅲ型(T3SEs)、Ⅳ型(T4SEs)、Ⅵ型(T6SEs)分泌效应蛋白[3]. 作为介导宿主细胞信号转导的关键分子,细菌效应蛋白(Effector proteins)的输入可使宿主细胞功能发生紊乱,以便细菌在宿主体内更好的生存、繁殖与感染,故效应蛋白在病菌与宿主相互作用机制研究中扮演着重要角色.

鉴于细菌效应蛋白重要的生物学意义,研究人员提出了多种可准确识别细菌效应蛋白的预测方法,但它们大都只能识别某一类分泌效应蛋白,如 T3SEs[4-9], T4SEs[10-15], T6SEs[16-18]. 在这三类分泌效应蛋白中,由于 T3SEs、T4SEs 均不含 N 端信号肽,且二者可能具有相似的进化保守性或序列模体(Motifs)[10],故现有计算方法极难区分这两类效应蛋白[19]. 为了解决这一问题,基于支持向量机(SVM)算法和伪位置特异性得分矩阵(PsePSSM),本文构建了一个二元分类器以快速准确地区分革兰氏阴性菌Ⅲ型、Ⅳ型分泌效应蛋白. 本方法对测试集总的预测准确率为 82.76%,表明其对 T3SEs 和 T4SEs 具有较好的区分能力,可作为一种辅助工具用于分泌效应蛋白在病原菌-宿主相互作用分子机制方面的研究.

2 材料与方法

2.1 材 料

本文从细菌分泌效应蛋白数据库(SecretEPDB)[20]中得到了实验所需的大部分数据. SecretEPDB 收录了 T3SEs、T4SEs、T6SEs 三类分泌效应蛋白的相关数据,并提供了蛋白质的特征、功能、二级结构、Pfam 结构域、代谢途径、进化细节等信息. 通过该数据库,共收集得到 1 230 条 T3SEs 和 731 条 T4SEs. 此外,我们从文献[4]和[13]中分别得到 35 条 T3SEs 和 30 条 T4SEs. 移除重复序列(即训练集或测试集中已有蛋白质序列)后,独立测试集中这两类效应蛋白各剩 25 条.

2.2 建模方法

分泌效应蛋白预测作为一种常见的蛋白质分类问题,已有越来越多的机器学习算法参与其中,如支持向量机(SVM)[8, 10-12, 14-15, 17]、隐马尔可夫模型(HMM)[5-6, 18]、随机森林(RF)[4]、深度学习(DL)[9]等. 在这些机器学习算法中,SVM 是应用最广泛的算法[3]. 此外,由于 SVM 在前期革兰氏阴性菌分泌蛋白的分类研究中[2, 19]已有成功的应用,故本文也选取 SVM 来构建预测模型.

2.3 模型的性能评估参数

本文中,灵敏度(SE),特异性(SP),准确率(ACC)和马氏相关系数(MCC)[21]分别被用于模型预测能力的评估.

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

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

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

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

其中,TP 为真阳性,即正样本的准确识别数;FP 表示假阳性,即负样本的错误识别数;TN 表示真阴性,即负样本的准确识别数;FN 表示假阴性,即正样本的错误识别数.

3 实验部分

3.1 实验数据

为去除实验数据中相似的蛋白质序列,增强预测模型的稳健性,采用 CD-HIT Suite[22]对原始数据进行处理后(序列相似度阈值 25%),得到 302 条 T3SEs 和 375 条 T4SEs. 通过 MATLAB 工具箱对序列随机后,选取其中的 70%作为训练集,其

表 1 本文所用实验数据集

Tab. 1 All experimental data sets used in this article

类别	原始数据	去冗余后数据	训练集	测试集	独立测试集
T3SEs	1 230	302	211	91	25
T4SEs	731	375	263	112	25
合计	1 961	677	474	203	50

余 30%作为测试集[23]. 结合 2.1 节所述的独立测试集，本文所用实验数据集均列于表 1 中.

3.2 特征提取与替代模型

不同类型的分泌效应蛋白，通常在序列、结构、功能等方面存在一定差异. 为准确区分 T3SEs 与 T4SEs，本文分别采用氨基酸组成、位置特异性得分矩阵、自协方差变量以表征蛋白质序列中氨基酸残基的频率信息、进化信息及邻接效应.

氨基酸组成(AAC)常用于表征 20 种天然氨基酸在蛋白质序列中出现的频率信息，每条蛋白质均被转化为一个 20 维的数字向量.

进化信息在蛋白质的分类研究中发挥着越来越重要的作用，而位置特异性得分矩阵(PSSM)则常用于表征蛋白质序列中氨基酸的进化信息[24]. 以期望值阈值为 10^{-3}，通过 PSI-BLAST 程序搜索 Swiss-Prot 数据库，经 3 次迭代后，可得到每条蛋白质的位置特异性得分矩阵. 通过相关计算公式[25]对这些矩阵进行转换后，每条蛋白质均被表征为一个 20 维的数字向量.

为有效表征蛋白质序列中氨基酸残基间的相互作用关系，自协方差(AC)变量常用于计算残基间的邻接效应. 自协方差(AC)变量的有关计算公式已详细描述于相关论文中[25]，故本文不再赘述. 经自协方差变换后，每条蛋白质均被转换为一个 25 维的向量.

基于 AAC、PSSM 和 AC，我们共构建了 4 个蛋白质替代模型：模型 1 仅含 AAC；模型 2 仅含 PSSM；模型 3 为 AAC 与 AC 合并而成的伪氨基酸组成(PseAAC)；模型 4 为 AAC 与 PSSM 合并而成的伪位置特异性得分矩阵(PsePSSM).

3.3 模型的构建

本文通过 libsvm-3.22 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)工具箱构建了最终的 SVM 预测模型. 模型核函数为径向基函数(RBF)，且通过网格搜索法对其正则化参数 C 和宽度参数 γ 进行优化. 虽然目前已有多种交叉验证方法被用于统计预测中，留一法(Leave-one-out)被认为是最客观公正的[26]，故本研究也采用留一法建立了最终的预测模型.

4 结果与讨论

4.1 替代模型的确定

根据 3.2 节描述的 4 个蛋白质替代模型，我们构建了 4 个 SVM 预测模型，它们对训练集的测试结果均列于表 2 中.

表 2　不同替代模型对训练结果的影响

Tab. 2　Performance of different substitution models

模型	C	γ	准确率/%
模型 1	512	0.007 812 5	82.489 5
模型 2	8.0	0.031 25	79.113 9
模型 3	2.0	0.125	83.122 4
模型 4	8.0	0.5	85.654 0

由表 2 可看出，模型 2 的训练效果最差，表明 T3SEs 和 T4SEs 在序列进化保守性上的确可能存在一定的关联性. 模型 3、模型 4 与模型 1、模型 2 的训练结果表明，替代模型中所含特征越多，其包含的信息量就越大，模型的预测性能也越强. 此外，模型 4 的训练结果优于模型 3 的，表明 PSSM 所包含的信息量可能多于 AC 的. 由于模型 4 的训练结果最好，且其核函数参数也较为合理，故本文拟选择该模型作为最终的蛋白质替代模型.

4.2 模型的实际应用

测试集数据首先被用于模型 3 与模型 4 实际预测性能的进一步比较，相关测试结果均列于表 3 中.

表 3　不同 SVM 模型对测试集的预测结果

Tab. 3　Prediction results of different SVM models obtained by analyzing the test sets

类别	T3SEs	T4SEs	合计
测试集数据	91	112	203
模型 3			
准确预测数	77	89	166
准确率/%	84.62	79.46	81.77
模型 4			
准确预测数	76	92	168
准确率/%	83.52	82.14	82.76

如表 3 所示，模型 4 准确识别出测试集中 76 条 T3SEs 和 92 条 T4SEs，其对这两类效应蛋白的预测准确率均超过 80%，且总的准确率为 82.76%，略优于模型 3 的 81.77%，表明将模型 4 作为最终的预测模型是正确的.

根据不同方法间交叉验证测试结果[3]，BEAN 2.0 对 T3SEs 的预测性能最好，而 T4Effpred 则被认为是 T4SEs 预测的最佳工具. 利用 2.1 节构建的独立测试集，我们进一步探讨了本方法、BEAN

2.0 及 T4Effpred 对这两类分泌效应蛋白的预测性能,相关测试结果如表 4 所示.

表 4 三种方法对独立测试集的预测结果

Tab. 4 Prediction results of the three methods obtained by analyzing the independent test sets

类别	T3SEs	T4SEs	合计
测试集数据	25	25	50
本方法			
准确预测数	22	16	38
准确率/%	88	64	76
BEAN 2.0			
准确预测数	23	—	—
准确率/%	92	—	—
T4Effpred			
准确预测数	—	7	—
准确率/%	—	28	—

由表 4 可看出,本方法准确识别出独立测试集中 22 条 T3SEs 和 16 条 T4SEs,总的预测准确率为 76%. 作为 T3SEs 的专业预测软件,BEAN 2.0 准确识别出 23 条 T3SEs,预测准确率高达 92%,但 25 条 T4SEs 有 3 条被错误预测为 T3SEs. T4Effpred 仅准确识别出 25 条 T4SEs 中的 7 条,预测准确率仅为 28%,且 25 条 T3SEs 中有 10 条被错误预测为 T4SEs. 这些实验结果再一次表明,T3SEs 与 T4SEs 的确可能具有相似的序列模体和进化保守性,故两者之间难以完全区分. 此外,虽然本方法对 T3SEs、T4SEs 的区分能力仍不是特别理想,但从整体上看是较为准确可靠的.

5 总 结

分泌效应蛋白重要的生物学意义推动了相关计算方法的开发,而这些计算方法的快速发展又反过来促进了对宿主与病原体间相互作用、细菌感染与毒力特性等方面的深入研究. 基于支持向量机和伪位置特异性得分矩阵,本文构建了一个可快速准确区分 T3SEs 与 T4SEs 的二元分类预测器. 实验结果表明,本方法对革兰氏阴性菌Ⅲ型、Ⅳ型分泌效应蛋白具有较强的区分能力,可作为辅助工具用于分泌效应蛋白的进一步研究. 此外,实现对 T6SEs 的准确预测仍是一项具有挑战性的任务,这也为我们下一步的研究指明了方向.

参考文献:

[1] Wandersman C. Concluding remarks on the special issue dedicated to bacterial secretion systems: function and structural biology [J]. Res Microbiol, 2013, 164: 683.

[2] Yu L, Liu F, Du L, *et al*. An improved approach for rapidly identifying different types of Gram-negative bacterial secreted proteins [J]. Nat Sci, 2018, 10: 168.

[3] An Y, Wang J, Li C, *et al*. Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems Ⅲ, Ⅳ and Ⅵ [J]. Brief Bioinform, 2018, 19: 148.

[4] Yang X, Guo Y, Luo J, *et al*. Effective identification of Gram-negative bacterial type Ⅲ secreted effectors using position-specific residue conservation profiles [J]. PLoS One, 2013, 8: e84439.

[5] Dong X, Zhang Y J, Zhang Z. Using weakly conserved motifs hidden in secretion signals to identify type-Ⅲ effectors from bacterial pathogen genomes [J]. PLoS One, 2013, 8: e56632.

[6] Yang Y, Qi S. A new feature selection method for computational prediction of type Ⅲ secreted effectors [J]. Int J Data Min Bioinform, 2014, 10: 440.

[7] Hobbs C K, Porter V L, Stow M L S, *et al*. Computational approach to predict species-specific type Ⅲ secretion system (T3SS) effectors using single and multiple genomes [J]. BMC Genomics, 2016, 17: 1048.

[8] Wang J, Li J, Yang B, *et al*. Bastion3: a two-layer ensemble predictor of type Ⅲ secreted effectors [J]. Bioinformatics, 2019, 35: 2017.

[9] Xue L, Tang B, Chen W, *et al*. DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type Ⅲ secreted effectors using the N-terminal sequence [J]. Bioinformatics, 2019, 35: 2051.

[10] Zou L, Nan C, Hu F. Accurate prediction of bacterial type Ⅳ secreted effectors using amino acid composition and PSSM profiles [J]. Bioinformatics, 2013, 29: 3135.

[11] Wang Y, Wei X, Bao H, *et al*. Prediction of bacterial type Ⅳ secreted effectors by C-terminal features [J]. BMC Genomics, 2014, 15: 50.

[12] Wang Y, Guo Y, Pu X, *et al*. Effective prediction of bacterial type Ⅳ secreted effectors by combined features of both C-termini and N-termini [J]. J

Comput Aided Mol Des，2017，31：1029.

[13] Wang J，Yang B，An Y，*et al*. Systematic analysis and prediction of type Ⅳ secreted effector proteins by machine learning approaches [J]. Brief Bioinform，2019，20：931.

[14] Xiong Y，Wang Q，Yang J，*et al*. PredT4SE-stack：prediction of bacterial type Ⅳ secreted effectors from protein sequences using a stacked ensemble method [J]. Front Microbiol，2018，9：2571.

[15] Ashari Z E，Brayton K A，Broschat S L. Using an optimal set of features with a machine learning-based approach to predict effector proteins for *Legionella pneumophila* [J]. PLoS One，2019，14：e0202312.

[16] Liang X，Moore R，Wilton M，*et al*. Identification of divergent type Ⅵ secretion effectors using a conserved chaperone domain [J]. P Natl Acad Sci USA，2015，112：9106.

[17] Wang J，Yang B，Leier A，*et al*. Bastion6：a bioinformatics approach for accurate prediction of type Ⅵ secreted effectors [J]. Bioinformatics，2018，34：2546.

[18] Nguyen T T，Lee H H，Park I，*et al*. Genome-wide analysis of type Ⅵ system clusters and effectors in *Burkholderia* species [J]. Plant Pathol J，2018，34：11.

[19] Yu L，Luo J，Guo Y，*et al*. In silico identification of Gram-negative bacterial secreted proteins from primary sequence [J]. Comput Biol Med，2013，43：1177.

[20] An Y，Wang J，Li C，*et al*. SecretEPDB：a comprehensive web-based resource for secreted effector proteins of the bacterial types Ⅲ，Ⅳ and Ⅵ secretion systems [J]. Sci Rep，2017，7：41031.

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

[22] Huang Y，Niu B，Gao Y，*et al*. CD-HIT suite：a web server for clustering and comparing biological sequences [J]. Bioinformatics，2010，26：680.

[23] Jiang J，Chen Y，Narayan A. Offline-enhanced reduced basis method through adaptive construction of the surrogate training set [J]. J Sci Comput，2017，73：853.

[24] Yu B，Li S，Qiu W，*et al*. Prediction of subcellular location of apoptosis proteins by incorporating PseP-SSM and DCCA coefficient based on LFDA dimensionality reduction [J]. BMC Genomics，2018，19：478.

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

[26] Huang G，Zhang Y，Chen L，*et al*. Prediction of multi-type membrane proteins in human by an integrated approach [J]. PLoS One，2014，9：e93553.

引用本文格式：

中　文：柳凤娟，杨庆，陈倩，等. 一种快速准确区分Ⅲ型、Ⅳ型分泌效应蛋白的计算方法[J]. 四川大学学报：自然科学版，2020，57：781.

英　文：Liu F J，Yang Q，Chen Q，*et al*. A fast and accurate computational approach for the distinction of type Ⅲ and Ⅳ secreted effector proteins [J]. J Sichuan Univ：Nat Sci Ed，2020，57：781.