Biophysics 生物物理学, 2017, 5(2), 9-15
Published Online May 2017 in Hans. http://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2017.52002

Hans汉斯

Addition of Protein Secondary Structure Information for Prediction of Anticancer Peptide

Wei Zhao, Yonge Feng*

College of Sciences, Inner Mongolia Agricultural University, Hohhot Inner Mongolia
Email: *fengyonge@163.com

Received: May 5th, 2017; accepted: May 23rd, 2017; published: May 26th, 2017

Abstract

The anticancer peptide is a kind of antimicrobial peptide which has obvious antitumor activity. Anti-cancer peptide can not only quickly and effectively eliminate pathogenic bacteria, but also can be effective in human tumor cells. Based on the published literature anticancer peptide dataset, the three kinds of protein secondary structure (3*PSS*) were extracted as the characteristic parameters for the first time. Combined with 20 kinds of amino acids (20*AAC*) and 6 kinds of hydrophobic amino acids (6*HP*) as characteristic information, using the quadratic discriminant method (*QD*) to carry out prediction, under the 7 folds of cross-validations, when using three kinds of protein secondary structure components (3*PSS*) combined with six kinds of hydrophobic amino acid components (6*HP*) as a feature, the overall accuracy (*Acc*) was 86%. When using three kinds of protein secondary structure components (3*PSS*) combined with 20 kinds of amino acid components (20*AAC*) as a feature, the total accuracy was 94%. It is found from the prediction results that the total accuracy can be improved with the addition of secondary structure information. In addition, compared with other prediction software, we confirm that the prediction accuracy was the highest when we join the secondary structure.

Keywords

Anticancer Peptide, 7-Folds of Cross-Validations, Protein Secondary Structure, Quadratic Discriminant Analysis (*QD*)

添加二级结构作为特征来对抗癌肽进行预测

赵　微，冯永娥*

内蒙古农业大学理学院，内蒙古　呼和浩特
Email: *fengyonge@163.com

*通讯作者。

文章引用: 赵微, 冯永娥. 添加二级结构作为特征来对抗癌肽进行预测[J]. 生物物理学, 2017, 5(2): 9-15.
https://doi.org/10.12677/biphy.2017.52002

收稿日期：2017年5月5日；录用日期：2017年5月23日；发布日期：2017年5月26日

摘　要

抗癌肽是一种具有明显抗肿瘤活性的抗微生物肽，抗癌肽不仅能快速高效地消灭致病病菌，还能有效地作用于人体肿瘤细胞。本文将已发表文献的抗癌肽数据集中，首次提取了蛋白质3种二级结构组分(3*PSS*)作为特征参量，并结合20种氨基酸组分(20*AAC*)和6种亲疏水氨基酸组分(6*HP*)作为特征信息，并采用二次判别法(*QD*)实施预测。最后，在7折交叉检验下，当采用蛋白质3种二级结构组分(3*PSS*)结合6种亲疏水氨基酸组分(6*HP*)作为特征时，预测总精度(*Acc*)达到86%；当采用蛋白质3种二级结构组分(3*PSS*)结合20种氨基酸组分(20*AAC*)作为特征时，预测总精度达到94%。从预测结果发现：添加了二级结构信息后，预测精度都有不同程度的提高。另外，在同种数据集中，和其他预测软件相比较，再次确认了加入二级结构信息作为特征后，我们的预测精度是最高的。

关键词

抗癌肽，7折交叉检验，蛋白质二级结构，二次判别法

1. 引言

抗癌肽(anticancer peptides，*ACP*)是一种具有明显抗肿瘤活性的抗微生物肽[1]，近年来生物学家们在抗癌肽这一领域做了大量研究，随着研究的深入，科学家们惊奇地发现，抗癌肽不仅能快速高效地消灭致病病菌，还能有效地作用于人体肿瘤细胞，使肿瘤细胞的核染色体合成受阻，造成 *DNA* 断裂而导致肿瘤细胞死亡[2] [3] [4]，抗癌肽(*ACP*)的发现为癌症治疗提供了新的希望[5] [6]，因为抗癌肽(*ACP*)不会损害正常的机体生理功能。在过去十年中，许多针对各种肿瘤类型的抗癌肽已经在临床上应用[7] [8] [9] [10]，表明抗癌肽(*ACP*)可能成为癌症治疗的一种手段。

2013 年，Tyagi 等人[11]基于氨基酸组分等特征信息，运用支持向量机(Suppoa Vector Machine，*SVM*)算法对抗癌肽进行预测，总精度(*Acc*)为 88.89%；2014 年，Hajisharifi 等人[12]基于伪氨基酸组分信息和局部比对内核理论，运用支持向量机(Suppoa Vector Machine，*SVM*)算法在 5 折交叉检验下对抗癌肽进行了预测，总精度(*Acc*)达到 89.7%；2016 年 Chen 等人[13]基于序列工具在五折交叉检验下对抗癌肽进行预测，总精度(*Acc*)达到 94.77%。

本文应用二次判别法[14] [15] [16]，选取 20 种氨基酸组分(20*AAC*)、蛋白质 3 种二级结构组分(3*PSS*)信息和 6 种亲疏水氨基酸组分(6*HP*)作为特征参量进行预测，最好的预测总精度(*Acc*)达到 94%，同时在和其他预测算法进行对比时，结果显示应用二次判别法预测要优于其他预测算法。

2. 材料与方法

2.1. 资料库

为了便于和同类工作比较，本文也选取了 Hajisharifi 等人[12] (http://aps.unmc.edu/AP/main.php)构建

的抗癌肽数据集，称为数据集 P。其中正集包含 138 条抗癌肽序列，称为 P^{AC}；负集包含 206 条非抗癌肽序列，称为 $P^{non\text{-}AC}$。

2.2. 特征的选取

2.2.1. 蛋白质二级结构组分

本文选取蛋白质 3 种二级结构组分(3*PSS*)信息作为特征参量，蛋白质 3 种二级结构分别为：α 螺旋、β 折叠和无规则卷曲(*Coil*)。数据集中，正集 138 条抗癌肽序列，负集 206 条抗癌肽序列，其二级结构信息是由 *PSIPERD* [17]软件预测获得。

2.2.2. 氨基酸组分

本文选取的 20 种氨基酸组分(20*AAC*)信息作为特征参量。

2.2.3. 亲疏水氨基酸组分

本文中根据氨基酸的亲疏水[18]将 20 种氨基酸分为 6 大类，具体分类信息如下：将强亲水类氨基酸天冬氨酸(*D*)、精氨酸(*R*)、谷氨酰胺(*E*)、天冬酰胺(*N*)、谷氨酸(*Q*)、赖氨酸(*K*)、组氨酸(*H*)归为一类，记作 *H*；将强疏水类氨基酸丙氨酸(*A*)、蛋氨酸(*M*)、苯丙氨酸(*F*)、亮氨酸(*L*)、异亮氨酸(*I*)、缬氨酸(*V*)归为一类，记作 *L*；将弱亲水性或弱疏水性氨基酸丝氨酸(*S*)、苏氨酸(*T*)、酪氨酸(*Y*)、色氨酸(*Z*)归为一类，记作 *W*；剩余的 3 种氨基酸，即脯氨酸(*P*)、甘氨酸(*G*)、半胱氨酸(*C*)因其特殊的化学结构各成一类。这样 20 种氨基酸可以归并为 6 种(即 H，L，W，P，C，G)，我们计算了这 6 种亲疏水氨基酸组分(6*HP*)作为特征。

2.3. 二次判别法(*QD*)

早期的二次判别法(*QD*)是从正负两个集合中判断待测序列的归属，本文参考冯永娥和罗辽复[14] [15] [16]的工作，使用广义的二次判别法，对任意两个集合 i，j 之间的判定关系，用下面的公式(1)表示：

$$\xi_{ij} = \ln p(\omega_i \mid x) - \ln p(\omega_j \mid x) \tag{1}$$

根据贝叶斯理论，可以导出：

$$\begin{aligned}\xi_{ij} &= \ln\frac{p_i}{p_j} - \frac{\delta_i - \delta_j}{2} - \frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|} \\ &= \left(\ln p_i - \frac{1}{2}\delta_i - \frac{1}{2}\ln|\Sigma_i|\right) - \left(\ln p_i - \frac{1}{2}\delta_j - \frac{1}{2}\ln|\Sigma_j|\right)\end{aligned} \tag{2}$$

其中 P_i 和 P_j 分别为 i， j 两个集合的样本数

我们设定：

$$\eta_v = \ln p_v - \frac{\delta_v}{2} - \frac{1}{2}\ln|\Sigma_v| \tag{3}$$

$$\delta_v = (R - \mu_i)^{\mathrm{T}} \Sigma_v^{-1} (R - \mu_i) \tag{4}$$

平均值向量 μ_v 可用下式表示

$$\mu_i = \frac{1}{p_v}\sum_{i=1}^{p_v} t_i \tag{5}$$

其中 p_v 是某一集合中的序列总数， δ_v 是 R 和 μ_v 之间的马氏距离。$|\Sigma_v|$ 是协方差矩阵 Σ_v 的行列式。(注意： μ_v 和 $|\Sigma_v|$ 在训练集中给定。)

$\sum_v$ 为协方差矩阵：

$$\sum_v = \begin{bmatrix} \sigma_{1,1}^v & \sigma_{1,2}^v & \cdots & \sigma_{1,6}^v \\ \sigma_{2,1}^v & \sigma_{2,2}^v & \cdots & \sigma_{2,6}^v \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{6,1}^v & \sigma_{6,2}^v & \cdots & \sigma_{6,6}^v \end{bmatrix}$$

$\sum_v$ 协方差矩阵的元素：

$$\sigma_{i,j}^v = \frac{1}{p_v}\sum\left(t_i - \mu_i^{(v)}\right)\left(t_j - \mu_j^{(v)}\right) \quad \left(v = P^{AC} \text{ and } P^{non\text{-}AC}\right) \tag{6}$$

在方程(2)中令 ξ_{ij} 写成 η_i，η_j 两数之差：

$$\xi_{ij} = \eta_i - \eta_j \tag{7}$$

我们可以从 η_i，η_j 的大小来判断出 $p\left(\omega_k \mid X\right)$ $\left(v = P^{AC} \text{ and } P^{non\text{-}AC}\right)$的大小顺序。即对于一个待测蛋白质序列 X，如果分别算出 η_i，η_j 然后进行比较这 2 个 η 的大小，如果 η_k 是 η_v 中最大的一个，很容易证明 $p\left(\omega_k \mid X\right)$ 就是 $p\left(\omega_v \mid X\right)$ 中最大的，说明 X 出现在这个类别中的概率最大，则待测序列 X 就属于可 k 类。

然而在统计算法中，经常伴随着涨落现象，所以我们定义了允许误差范围内的修正系数：

$$R = \frac{\eta_{corr} - \eta_{wro}}{\eta_{corr}} \tag{8}$$

其中，η_{corr} 表示属于自己结构的 η，η_{wro} 表示被错误预测为其他结构的 η (属于某类，但被预测为此类的 η)。我们通过对 R 值进行合适的设定，就能利用公式(8)对预测结果进行适当的修正。

2.4. 随机森林算法(*RF*)

随机森林[19] [20]算法是 *Leo Breiman* 在 2001 年提出的一种分类预测模型，是由许多单棵分类回归树组合而成的，一棵分类回归树就是一个分类器，最后的决策结果由投票法决定。它的基本思想是将很多弱分类器集成一个强分类器。随机森林算法是一种通过自助法重采样来构造多个分类器的组合分类器。

随机森林有两个重要的参数，一个是单棵决策树每个节点处分裂时所选用的候选特征参数的个数 m，另一个是随机森林中决策树的棵数 k (k = 500)。用随机森林分类器对新的数据进行判别与分类，按照树分类器进行投票，最后由投票法决定分类结果。随机森林通过在每个节点处随机选择特征进行分支，这样可以最小化各棵分类树之间的相关性，提高分类的精确性。本文使用的是 *R* 语言下的 *RF* 程序包。随机森林算法不会出现过度拟合现象、分类效率也很高，而且能够快速处理大样本数据，同时需要调整的参数也比较少，能更好的估计哪个特征在分类中更重要。

2.5. 分类预测性能评估

目前，预测算法性能检验常用的方法主要有独立检验(*independent test*)和 *K*-折交叉检验(*k-fold cross-validation test*)。

本文采用 7 折交叉检验，即将数据集随机分为 7 个子集合，依次从中取出一个子集作为测试集，而将剩余的 6 个子集合则作为训练集，此过程一共循环 7 次。对于任何预测算法性能的评价，主要是保证该预测算法能对属于同一数据域的新样本具有推广性能[21]。在我们的研究中主要使用敏感性(*Sn*)、特异性(*Sp*)、预测总精度(*Acc*)和 *Mathew's* 相关系数 *MCC* 这 4 个指标评价预测算法的有效性：

1) 敏感性：表示数据集中每一类的预测正确率，定义如下：

$$Sn = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

2) 特异性：表示分类预测中每个类别预测结果的可信度，定义如下：

$$Sp = \frac{TN}{TN + FP} \times 100\% \tag{10}$$

3) 预测总精度：表示预测的总体正确率，定义如下：

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \tag{11}$$

4) 马修相关系数：一个整体评价指标，反应预测的综合能力，定义如下：

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \times 100\% \tag{12}$$

其中，*TP* 表示该类中正确预测的样品数，*FN* 表示该类中错误预测的样品数，*FP* 表示其他类被预测为此类的样品数，*TN* 表示其他类中正确预测为其他类的样品数，*N* 表示样品总数，敏感性(*Sn*)表示预测算法的能力，特异性(*Sp*)表示结果预测的可信度，预测总精度(*Acc*)表示预测结果的正确率，Mathew's 相关系数 *MCC* 表示对预测算法的综合评价，*MCC* 的取值范围在[−1，+1]之间，*MCC* 值等于 1 时，表示预测结果与真实类别完全相关，*MCC* 值等于 0 时，表示是完全随机的预测，*MCC* 值等于-1 时，表示负相关性。

3. 结果与讨论

3.1. 二次判别法(QD)对抗癌肽的预测

在抗癌肽数据集 *P* 中，我们分别提取 20 种氨基酸组分(20*AAC*)、蛋白质 3 种二级结构组分(3*PSS*)、6 种亲疏水氨基酸组分(6*HP*)作为特征参量，应用二次判别法进行预测，在 7 折交叉检验下，取 $R < 0.3$，预测结果列在表 1。

由表 1 可见，在 7 折交叉验证下，采用二次判别法，选取 20 种氨基酸组分(20*AAC*)结合蛋白质 3 种二级结构组分(3*PSS*)信息作为特征参量时，预测正确率最高，预测总精度(*Acc*)达到 94%，敏感性 *Sn*、特异性 *Sp* 与马修相关系数 *MCC* 分别为 90%、96.67%、和 0.87。由表 1 可见：在序列信息作为特征参量的基础上，加入了二级结构信息，预测精度都有了很大的提高，说明蛋白质二级结构信息是非常有效的特征参量。

3.2. 随机森林(*RF*)算法对抗癌肽的预测

为了确定我们所选的特征参量{20 种氨基酸组分(20*AAC*)、蛋白质 3 种二级结构(3*PSS*)、6 种亲疏水氨基酸组分(6*HP*)}对于识别抗癌肽具有不错的效果，我们又使用了随机森林算法结合这些特征参数实施预测，7 折交叉检验，取 $R < 0.3$，预测结果列在表 2。

通过表 2，再次确认加入了二级结构信息，预测精度都有了不错的提高。最后，在 7 折交叉检验下，采用随机森林算法，选取 20 种氨基酸组分(20*AAC*)结合蛋白质 3 种二级结构组分(3*PSS*)信息作为特征参量，预测正确率最高，总正确率(*Acc*)最高达到 88%，敏感性 *Sn*、特异性 *Sp* 与马修相关系数 *MCC* 分别为 80%、93%、和 0.75。同时对比表 1 的结果，发现在相同的特征参数下，二次判别法(*QD*)比随机森林(*RF*)更适合于抗癌肽的预测。

近几年，研究者对于抗癌肽作了大量的研究，为了显示我们预测模型的有效性，在相同的抗癌肽数据集 *P* 下，我们对比了 *Hajisharifi* [12]等人的研究结果，对比结果列在表 3。

Table 1. Prediction results based on *QD* in 7 fold cross-validation
表 1. 7 折交叉下二次判别法的预测结果

Features	Sn (%)	Sp (%)	Acc (%)	MCC
3*PSS*	35	96.67	72	0.43
6*HP*	95	76.67	84	0.70
3*PSS* + 6*HP*	85	86.67	86	0.72
20*AAC*	80	93.33	92	0.83
20*AAC* + 3*PSS*	90	96.67	94	0.87

Notes: *PSS*：蛋白质二级结构；*AAC*：氨基酸组份；*HP*：亲疏水氨基酸

Table 2. Prediction results based on *RF* in 7- fold cross-validation
表 2. 7-折交叉下随机森林法的预测结果

Features	Sn (%)	Sp (%)	Acc (%)	MCC
3*PSS*	50	80	68	0.32
6*HP*	70	87	72	0.41
3*PSS* + 6*HP*	60	80	80	0.52
20*AAC*	80	90	86	0.71
20*AAC* + 3*PSS*	80	93	88	0.75

Notes: *PSS*：蛋白质二级结构；*AAC*：氨基酸组份；*HP*：亲疏水氨基酸

Table 3. Comparison results by using different method
表 3. 不同方法的对比结果

Methods	Sn (%)	Sp (%)	Acc (%)	MCC
Our method	90	96.67	94	0.87
Hajisharifi [12]	81.48	85.36	83.82	0.66

Notes: *Hajisharifi* [12] see in ref [12].

表 3 的对比结果显示了我们的结果是优于其它工作的，可见我们的预测模型值得推广。尤其是加入结构信息后，预测精度有很大的提高，说明添加二级结构信息在抗癌肽预测中是个不错的选择。

4. 结论

在本文中，我们首次将蛋白质 3 种二级结构组分(3*PSS*)作为特征，并结合 20 种氨基酸组分(20*AAC*)、6 种亲疏水氨基酸组分(6*HP*)，应用二次判别法(*QD*)进行预测，结果显示：应用 20 种氨基酸组分(20*AAC*)结合蛋白质 3 种二级结构组分(3*PSS*)进行预测时正确率较高，总精度(*Acc*)最高为 94%，并且高于其它的预测算法[19] [20]。希望我们的预测模型可以运用到其它抗微生物肽的识别中。

基金项目

本项目由国家自然基金项目(31360206)；内蒙古自治区高等学校科研项目(NJZY067)；内蒙古农业大学基础科研基金(JC2013004)资助。

参考文献 (References)

[1] Reddy, K.V.R., Yedery, R.D. and Aranha, C. (2004) Antimicrobial Peptides: Premises and Promises. *International Journal of Antimicrobial Agents*, **24**, 536-547. https://doi.org/10.1016/j.ijantimicag.2004.09.005

[2] Meng, S. and Wang, F.S. (201l) Research Progress of Antimicrobial Peptides with Anticancer Activities. *Chinese Journal of Biochemical Pharmaceutics*, **32**, 241-244.

[3] Shi, S.L., Wang, Y.Y., *et a1*. (2006) Effects of Tachypiesin and N-Sodium Butyrate on Proliferation and Gene Expression of Human Gastric Adenocarcinoma Cell Line BGC-823. *World Journal of Gastroe-Nterology*, **12**, 1694-1698. https://doi.org/10.3748/wjg.v12.i11.1694

[4] Mader, J.S., Smyth, D., Marshall, J., *et a1*. (2006) Bovine Lactoferricin Inhibits Basic Fibroblast Growth Factor-and Vascular Endothelial Growth Factor165-Induced Angiogenesis by Competing for Heparin-Like Binding Sites on Endothelial Cells.*The American Journal of Pathology*, **169**, 1753-1766. https://doi.org/10.2353/ajpath.2006.051229

[5] Gaspar, D., Veiga, A.S. and Castanho, M.A. (2013) From Antimicrobial to Anticancer Peptides. *Frontiers in Microbiology*, **4**, 1-16. https://doi.org/10.3389/fmicb.2013.00294

[6] Huang, Y., Feng, Q., Yan, Q., *et al*. (2015) Alpha-Helical Cationic Anticancer Peptides: A Promising Candidate for Novel Anticancer Drugs. *Mini Reviews in Medicinal Chemistry*, **15**, 73-81. https://doi.org/10.2174/1389557514666141107120954

[7] Hariharan, S., Gustafson, D., Holden, S., *et al*. (2007) Assessment of the Biological and Pharmacological Effects of the Alpha Nu Beta3 and Alpha Nu Beta5 Integrin Receptor Antagonist. *Annals of Oncology*, **18**, 1400-1407.

[8] Gregorc, V., De Braud, F.G., De, T.M., *et al*. (2011) A Selective Vascular Targeting Agent in Combination with Cisplatin in Refractory Solid Tumors. *Clinical Cancer Research*, **17**, 1964-1972. https://doi.org/10.1158/1078-0432.CCR-10-1376

[9] Sah, B.N.P., Asiljevic, T.V., McKechnie, S., *et al*. (2015) Identification of Anticancer Peptides from Bovine Milk Proteins and Their Potential Roles in Management of Cancer: A Critical Review. *Comprehensive Reviews in Food Science and Food Safety*, **2**, 123-138. https://doi.org/10.1111/1541-4337.12126

[10] kobon, T.E., Thongararm, P. and Roytrakul. S. (2016) Prediction of Anticancer Peptides against MCF-7 Breast Cancer Cells from the Peptidomes of Achatina Fulica Mucus Fractions. *Computational and Structural Biotechnology Journal*, **14**, 49-57.

[11] Tyagi, A., Kapoor, P., Kumar, R. *et al*. (2013) In Silico Models for Designing and Discovering Novel Anticancer Peptides. *Scientific Reports*, **3**, 1-8. https://doi.org/10.1038/srep02984

[12] Hajisharifi, Z., Piryaiee, M., Mohammad, B.M., *et a1*. (2014) Predicting Anticancer Peptides with Chou's Pseudo Amino Acid Composition and Investigating their Mutagenicity via Amestest. *Journal of Theoretical Biology*, **341**, 34-40.

[13] Wei, C., Hui, D., Feng, P.M., Lin, H. *et al*. (2016) IACP: A Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget*, **13**, 16895-16909.

[14] Feng, Y.E. and Luo, L.F. (2008) Use of Tetrapeptide Signals for Protein Secondary Structure Prediction. *Amino Acids*, **35**, 607-614. https://doi.org/10.1007/s00726-008-0089-7

[15] Feng, Y.E., Lin, H. and Luo, L.F. (2014) Prediction of Protein Secondary Structure Using Feature Selection and Analysis Approach. *Acta Biotheoretica*, **62**, 1-14. https://doi.org/10.1007/s10441-013-9203-7

[16] Feng, Y.E. (2014) Prediction of Four Kinds of Simple Supersecondary Structures in Protein by Using Chemical Shifts.*The Scientific World Journal*, **2014**, Article ID: 978503.

[17] Liam, J.M., Kevin, B. and David, T.J. (2000) The PSIPRED Protein Structure Prediction Server. *Bioinformatics*, **16**, 404-405.

[18] Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

[19] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

[20] Li, Z.C., Lai, Y.H., Chert, L.L., *et al*. (2012) Identification of human protein complexes from Local Sub-Graphs of Protein Interaction Network Based on Random Forest with Topological Structure Features. *Analytica Chimica Acta*, **718**, 32-41. https://doi.org/10.1016/j.aca.2011.12.069

[21] Yang, H. and Xu, H.M. (2015) Protein Subcellular Localization Prediction Based on Reduced Representation of Amino Acid and Statistical Characteristic. *Chinese Journal of Bioinformatics*, **2**, 103-110.