Hans Journal of Computational Biology 计算生物学, 2016, 6(3), 62-71
Published Online September 2016 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2016.63008

Hans汉斯

Prediction of Apoptosis Protein Subcellular Localization Based on Hybrid Feature Parameters

Jixian Xue, Yingli Chen*, Yuanyuan Zhai

School of Physical Science and Technology, Inner Mongolia University, Hohhot Inner Mongolia
Email: *stchenyl@imu.edu.cn

Received: Sep. 8th, 2016; accepted: Sep. 26th, 2016; published: Sep. 29th, 2016

Abstract

Studies have shown that sequence and structure characteristics of the mRNA have a certain relevance with subcellular localization of protein. In this article, it extracted two mRNA information of apoptosis proteins: the three reading frame 3-mer mRNA sequence frequency information and mRNA secondary structure-sequence mode information, and to construct feature vector which indicate mRNA and amino acid sequence with physicochemical properties, stickiness and evolutionary information of apoptosis proteins. Meanwhile, by using support vector machine algorithm, apoptosis proteins of four different subcellular localizations were predicted. The study found that the hybrid of mRNA and AAs information promoted prediction result, and the overall prediction access rate achieved 82.18% while 78.26% for independent test datasets by the Jackknife test. Prediction results show that sequence and structure characteristics of the mRNA contribute to prediction of the subcellular localization of apoptosis proteins.

Keywords

Apoptosis Protein, mRNA Secondary Structure, Subcellular Localization

多种信息融合的细胞凋亡蛋白质的亚细胞定位预测

薛济先，陈颖丽*，翟媛媛

*通讯作者。

文章引用: 薛济先, 陈颖丽, 翟媛媛. 多种信息融合的细胞凋亡蛋白质的亚细胞定位预测[J]. 计算生物学, 2016, 6(3): 62-71. http://dx.doi.org/10.12677/hjcb.2016.63008

内蒙古大学物理科学与技术学院，内蒙古 呼和浩特
Email: *stchenyl@imu.edu.cn

收稿日期：2016年9月8日；录用日期：2016年9月26日；发布日期：2016年9月29日

摘 要

研究表明mRNA的序列和结构特性与蛋白质的亚细胞定位有一定关系。本文提取了细胞凋亡蛋白质的两种mRNA信息：三阅读框3-mer mRNA序列频数信息、mRNA二级结构-序列模式信息，并结合细胞凋亡蛋白质的氨基酸物理化学性质、氨基酸黏性特征和进化信息，构成特征向量来表示mRNA和蛋白质序列，利用支持向量机算法，对四种不同亚细胞位置的细胞凋亡蛋白质进行预测。研究发现融合mRNA信息与氨基酸信息后预测效果更佳，在Jackknife检验下，预测总精度达到82.18%，且独立测试集预测总精度达到78.26%。结果表明，mRNA的序列和结构特性有助于细胞凋亡蛋白质的亚细胞定位预测。

关键词

细胞凋亡蛋白，mRNA二级结构，亚细胞定位

1. 前言

细胞凋亡蛋白质是一类有着特殊功能的蛋白质，在生物体的生长发育和维持体内平衡中扮演着重要的角色[1]。这些蛋白质对于了解细胞凋亡的过程和机理具有重要作用。细胞凋亡与许多疾病相关，如自身免疫疾病、肿瘤、神经退行性病变等[2]。细胞凋亡蛋白质的功能与其亚细胞位置紧密相关[3]，从生物信息学的角度预测蛋白质在细胞中的位置能更好地了解它们的功能。

mRNA 是 RNA 分子中的一大家族，它将 DNA 中的遗传信息传递到核糖体中，在核糖体上作为蛋白质的合成模板，决定肽链中氨基酸的排列顺序。研究表明 mRNA 的二级结构对研究其功能具有重要的作用[4]。本文将细胞凋亡蛋白质所对应的 mRNA 的二级结构信息挑选出来，并与蛋白质一级序列信息相结合，利用生物信息学的方法统计分析了结构和序列信息，将有利于更深刻地了解不同亚细胞位置中细胞凋亡蛋白质的特性。

本文以细胞凋亡蛋白质的 mRNA 序列，mRNA 二级结构，氨基酸序列作为研究对象，统计分析了三阅读框 3-mer mRNA 序列频数信息，mRNA 二级结构-序列模式信息，氨基酸的物理化学性质，氨基酸黏性和进化信息，用支持向量机的方法基于 Jackknife 检验和独立检验对不同亚细胞位置的细胞凋亡蛋白质进行预测。

2. 数据集

本文所采用的细胞凋亡蛋白质数据均来源于 Uniprot 数据库(Release 2015_12 http://www.Uniprot.org/)。根据关键词 apoptosis 挑选出 1128 个细胞凋亡蛋白质，其中多定位的蛋白质有 555 个，有明确唯一单一定位信息的蛋白质有 572 个。去掉蛋白质数量过少的亚细胞位置，在 RefSeq 数据库中查找到选定的单一定位的蛋白质的 mRNA 序列，同时去掉 mRNA 序列长度大于 10,000 nt 的蛋白质。最终采用的数据集共包含蛋白质 331 个，分别位于四个亚细胞位置：细胞核，细胞膜，细胞质和线粒体，见图 1。数据集中每个蛋白质都有对应的 mRNA 序列。

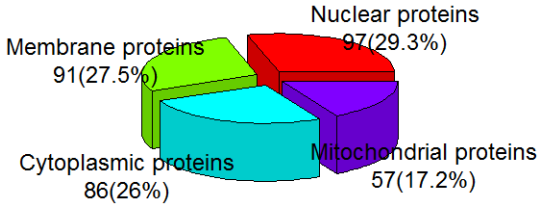

Figure 1. Apoptosis protein data distribution diagram
图 1. 凋亡蛋白质数据分布图

3. 特征选取

3.1. 三阅读框下 3-mer mRNA 序列信息

mRNA 由 DNA 转录而来，携带着重要的遗传信息，其中最基本的信息就是序列信息。mRNA 序列中包含了不同大小、形状和化学性质的四种碱基：胞嘧啶(C)、鸟嘌呤(G)、腺嘌呤(A)和尿嘧啶(U)，3-mer 序列信息是指序列中任意三个相邻的碱基。对于一条序列，由起始位点开始统计三联体，有三种阅读框(由 W1，W2，W3 表示)，取一个细胞凋亡蛋白质(图 1)对应的 mRNA 作为例子：

ATGTCGGGACCCGTGCCAAGCAGGGCCAGAGTTTACACAGATGTTAATACACACAGACCT.....

第一个阅读框(W1)，3-mer 序列信息：ATG TCG GGA CCC GTG CCA AGC AGG GCC .....

第二个阅读框(W2)，3-mer 序列信息：TGT CGG GAC CCG TGC CAA GCA GGG CCA .....

第三个阅读框(W3)，3-mer 序列信息：GTC GGG ACC CGT GCC AAG CAG GGC CAG .....

任一序列的三阅读框 3-mer 序列频数表示为公式(1) (2)

$$\text{3-mer} = [\text{W1}, \text{W2}, \text{W3}] \tag{1}$$

$$Wk = \left[x_1^k, x_2^k, \cdots, x_i^k \right] \left(k = 1,2,3; i = 1,2,\cdots,4^3 \right) \tag{2}$$

其中 x_i^k 表示第 k 个阅读框中第 i 个三联体出现的频数。

3.2. mRNA 二级结构 - 序列模式信息

研究表明，mRNA 的功能与其结构密切相关，为了使 mRNA 结构信息可以作为机器学习的特征参数，我们使用 RNAfold 软件[5]来预测 mRNA 的二级结构，二级结构的预测结果是以点或括号“ . 或 ( ) ”表示的。“( )”表示配对的碱基，形成茎结构；“.”表示不配对的碱基，形成单链或环结构。在本文中，为了计算方便对“(”和“)”不加以区分[4]。给出任一条 mRNA 序列，结构 - 序列模式信息可由二级结构预测图和相邻的三个碱基中中间的碱基表示。64 种三联体可约化为 32 种(4 × 23)组合，即：A(((、A((.、A(..、A...、A.((、A.(.、A..(、A.((、U(((、U((.、U(..、U...、U.((、U.(.、U..(、U.((、G(((、G((.、G(..、G...、G.((、G.(.、G..(、G.((、C(((、C((.、C(..、C...、C.((、C.(.、C..(、C.((。我们将这种三联体结构片段的频数作为结构 - 序列模式的特征参数。

3.3. 物理化学性质

AAIndex database [8]由 AAindex1、AAindex2 两部分组成，其中的 AAindex1 是氨基酸指数(amino acid indexes)表，这些指数是将氨基酸不同物理化学和生物学性质量化后的数据，目前共搜集了 544 种氨基酸指数。本文采用了其中九种物理化学性质[6] [7]，分别是 Hydrophilicity value (亲水性值)、Mean polarity (平均极性)、Isoelectric point (氨基酸等电点)、Refractivity (折射率)、Average flexibility indices (平均灵活性指标)、Average volume of buried residue (埋藏残基的平均体积)、Electron-ion interaction potential values (电子

离子相互作用势值)、Transfer free energy to surface (转换到表面的自由能)、Consensus normalized hydrophobicity (标准化后的疏水性)，利用同一蛋白质序列中不同距离的氨基酸残基之间存在的相互作用，获得更多的预测信息。九种物理化学性质见表 1，这些信息均来自 AAIndex database [8]。首先，用标准化后的氨基酸指数(amino acid indexes)表示序列中的每一个氨基酸残基，氨基酸指数同样提取于 AAIndex database，利用公式(3)对第 i 种氨基酸指数进行标准化。这种方式已被 Chou 等人用于物理化学性质的标准化[9]-[11]。

$$P_{normal\_i}^{(k)} = \frac{p_i^{(k)} - \overline{p}_i}{\sqrt{Var\left(p_i\right)}} \tag{3}$$

$$\overline{p} = \frac{1}{20}\sum_{k=1}^{20} p_i^{(k)} \tag{4}$$

$$Var\left(p_i\right) = \frac{1}{20}\sum_{k=1}^{20}\left(p_i^{(k)} - \overline{p}_i\right)^2 \tag{5}$$

标准化后，每一种特性变成了一组新的 20 个数字。对于第 i 种特性，任一条蛋白质序列可以表示为 $P_{(i)1}, P_{(i)2}, \cdots, P_{(i)L}$，其中 L 为序列长度，$P_{(i)k}$ $(1 \le k \le L)$是序列中第 k 个氨基酸残基的第 i 种特性的标准化值，然后利用等式(6)计算自相关函数的值。

$$R_i\left(\tau\right) = \frac{1}{L-\tau}\sum_{k=1}^{L-\tau} p_k^{(i)} p_{k+\tau}^{(i)} \tag{6}$$

$$R_i\left(\tau\right), 1 \le \tau \le T \tag{7}$$

其中，T 是一个常数。

通过计算得到的每一种特性的特征向量，都包含该条序列中不同距离的氨基酸残基间的相互作用关系。物理化学特性参数向量可表示为公式(8)：

$$V_i = \left[R_i\left(1\right), R_i\left(2\right), \cdots, R_i\left(T\right)\right]^{\mathrm{T}} \tag{8}$$

3.4. 氨基酸黏性信息 Stickiness of AAs

在生物体中存在着大量的蛋白质，其中功能性蛋白质的两两相互作用[12] [13]是很少的，而非功能性的、随机的蛋白质相互作用有很多。某些细胞特性有助于减少非功能性蛋白质的相互作用。基于蛋白质相互作用面和蛋白质表面溶剂可及性的氨基酸残基频数，氨基酸黏性(stickiness)可以被定义为[14]：

$$S = \log\left(\frac{f_{\mathrm{AA(interface)}}}{f_{\mathrm{AA(surface)}}}\right) \tag{9}$$

其中 $f_{\mathrm{AA(interface)}}$代表蛋白质相互作用界面(interface)的氨基酸频数，$f_{\mathrm{AA(surface)}}$代表蛋白质表面(surface)的氨基酸的频数。Levy 通过大量的分析发现蛋白质表面黏性的变化可能与它的亚细胞位置有关[14]。通过公式统计氨基酸黏性值，结果见表 2 [14]。

Petersen [15]等人开发了网站 Net Surf P (http:// www.cbs.dtu.dk/ ser vic es/NetSurfP-1.1/)，通过网站提交任务可预测蛋白质的表面位置信息，相对表面溶剂可及性(RSA)，绝对表面溶剂可及性，RSA 的 z-fit 值，α 螺旋概率，β 折叠概率和无规卷曲概率。利用这些预测结果，每条蛋白质序列可以构建成一个向量(10)：

Table 1. The 9 physicochemical properties
表 1. 9 种物理化学性质

Properties description	Reference
Hydrophilicity value	Hopp-Woods (1981)
Mean polarity	Radzicka-Wolfenden (1988)
Isoelectric point	Zimmerman *et al*. (1968)
Refractivity	McMeekin *et al*. (1964)
Average flexibility indices	Bhaskaran-Ponnuswamy (1988)
Average volume of buried residue	Chothia (1975)
Electron-ion interaction potential values	Cosic (1994)
Transfer free energy to surface	Bull-Breese (1974)
Consensus normalized hydrophobicity	Eisenberg (1984)

Table 2. The stickiness index of amino acids
表 2. 氨基酸黏性

amino acid (氨基酸)	stickiness index (氨基酸黏性)	amino acid (氨基酸)	stickiness index (氨基酸黏性)
A	0.0062	M	1.0124
C	1.0372	N	−0.2693
D	−0.7485	P	−0.1799
E	−0.7893	Q	−0.4114
F	1.2727	R	−0.0876
G	−0.1771	S	0.1376
H	0.1204	T	0.1031
I	1.1109	V	0.7599
K	−101806	W	0.7925
L	0.9138	Y	0.8806

$$P=\begin{bmatrix} A_1 & Sr_1 & Sa_1 & Z_1 \\ A_2 & Sr_2 & Sa_2 & Z_2 \\ \vdots & \vdots & \vdots & \vdots \\ A_j & Sr_i & Sa_i & Z_i \\ \vdots & \vdots & \vdots & \vdots \\ A_L & Sr_L & Sa_L & Z_L \end{bmatrix} \tag{10}$$

L 为序列长度，A 为蛋白质序列中的氨基酸残基，Sr 为 RSA，Sa 为绝对溶剂可及性，Z 为 RSA 的 z-fit 值。根据表 2，用氨基酸黏性值 S 代替公式(10)中的 A_j，得到向量(11)。

$$P=\begin{bmatrix} S_1 & Sr_1 & Sa_1 & Z_1 \\ S_2 & Sr_2 & Sa_2 & Z_2 \\ \vdots & \vdots & \vdots & \vdots \\ S_j & Sr_i & Sa_i & Z_i \\ \vdots & \vdots & \vdots & \vdots \\ S_L & Sr_L & Sa_L & Z_L \end{bmatrix} \tag{11}$$

然后利用自相关方程(12)计算各个预测信息，特征参数向量表示为公式(13)：

$$\Theta^{\Gamma}(\kappa)=\frac{1}{L-\kappa}\sum_{l=1}^{L-\kappa}\Gamma_1\Gamma_{l+\kappa}\ (\Gamma=S,Sr,Sa,Z,1\le\kappa<L) \tag{12}$$

$$P_{\Theta}^{\Gamma}=\left[\Theta^{\Gamma}(0),\Theta^{\Gamma}(1),\Theta^{\Gamma}(2),\Theta^{\Gamma}(3),\cdots,\Theta^{\Gamma}(\kappa);\right](\Gamma=S,Sr,Sa,Z,1\le\kappa<L) \tag{13}$$

3.5. 进化信息(PSSM)

本文通过本地运行PSI-BLAST [16]程序，用细胞凋亡蛋白质数据集中的每条序列与nr数据库(released on 04 2016)中的序列进行比对和评价。设置E-value值为0.001，经过三次迭代搜索，获得数据集中每条蛋白质序列的同源序列，构建位置特异性得分矩阵(position-specific scoring matrix, PSSM)，提取蛋白质进化的保守信息[17]。首先将矩阵以行为单位利用公式(14)进行标准化，

$$V_{i\to j}=\frac{V_{i\to j}^{0}-\overline{V}_i}{\sqrt{Var(V_i)}} \tag{14}$$

其中$V_{i\to j}^{0}$是由PSI-BLAST直接得到的得分，$\overline{V}_i$是二十种氨基酸的平均值，$\sqrt{Var(V_i)}$是标准差，L为序列长度。

对于一条序列长度为L的蛋白质P，标准化后的$PSSM$矩阵(15)可以表示为：

$$P_{PSSM}=\begin{bmatrix}V_{1\to1} & V_{1\to2} & \cdots & V_{1\to20}\\ V_{2\to1} & V_{2\to2} & \cdots & V_{2\to20}\\ \vdots & \vdots & & \vdots\\ V_{i\to1} & V_{i\to2} & \cdots & V_{i\to20}\\ \vdots & \vdots & & \vdots\\ V_{L\to1} & V_{L\to2} & \cdots & V_{L\to20}\end{bmatrix} \tag{15}$$

为了利用该序列顺序的信息，采用自相关函数(16)来得到特征参数向量(17)

$$\theta_i^{\lambda}=\frac{1}{L-\lambda}\sum_{j=1}^{L-\lambda}V_1V_{l+\kappa}\ \ (i=1,2,\cdots,20;1\le\lambda<L) \tag{16}$$

$$P_{PSSM}^{\lambda}=\left[\theta_1^{\lambda},\theta_2^{\lambda},\cdots,\theta_i^{\lambda},\cdots,\theta_{20}^{\lambda}\right] \tag{17}$$

4. SVM 算法

4.1. 支持向量机(Support Vector Machine, SVM)

支持向量机(Support Vector Machine, SVM)是一种用于解决分类和回归问题的机器分类算法。基本原理是将低维空间向量集映射到高维空间，通过选用适当的核函数和寻找最优分类面，使得不同类别样本之间的间隔最大化，从而有效地解决非线性分类问题。

本文使用支持向量机的C-SVC (C-Support Vector Classifier，C-支持向量分类器)类型，径向基核心函数，采用台湾大学林智仁(Lin Chih-Jen)教授开发的libsvm 3.21软件包[18]，通过搜寻最优C和γ值来训练细胞凋亡蛋白质中按照亚细胞位置类别作为标记的特征参数数据集，进行Jackknife检验[19] [20]和独立检验的预测。

4.2. 预测性能评估

本文采用了Jackknife检验和独立检验，Jackknife检验被认为是较严格和客观的统计检验方法；而独立检验则反映了算法对新序列的预测能力。在Jackknife检验中，假设细胞凋亡蛋白质数据集共有N条蛋

白质序列，把其中的每条蛋白质依次作为待测样本，剩下的 N-1 条蛋白质作为测试集测试，并给出这条细胞凋亡蛋白质的分类。在独立检验中，本文随机选取 80%的数据作为训练集，用 Jackknife 检验训练出模型，将其余 20%的数据作为测试集来检验模型的预测能力。本文采用的评价算法性能的指标为：敏感性 (Sensitivity, Sn)、特异性(Specificity, Sp)、预测成功率(Accuracy, Acc)、总体成功率(Overall accuracy, OA)和评价综合预测结果的相关性系数(Mathew's Correlation Coefficient, MCC)，定义如下：

$$\mathrm{Sn} = TP_i \big/ \left(TP_i + FN_i\right) \times 100\% \tag{18}$$

$$\mathrm{Sp} = TN_i \big/ \left(TN_i + FP_i\right) \times 100\% \tag{19}$$

$$\mathrm{ACC} = \left(TP_i + TN_i\right) \big/ \left(TP_i + FN_i + TN_i + FP_i\right) \times 100\% \tag{20}$$

$$\mathrm{OA} = \sum_i \left(TP_i + TN_i\right) \big/ \sum_i \left(TP_i + FN_i + TN_i + FP_i\right) \times 100\% \tag{21}$$

$$\mathrm{MCC} = \left(TP_i \times TN_i - FP_i \times FN_i\right) \big/ \sqrt{\left(TP_i + FN_i\right) \times \left(TN_i + FN_i\right) \times \left(TP_i + FP_i\right) \times \left(TN_i + FP_i\right)} \tag{22}$$

其中，TP_i 表示第 i 类亚细胞位置中预测正确的细胞凋亡蛋白质条数；TN_i 表示非第 i 类亚细胞位置中的细胞凋亡蛋白质被正确的识别为非 i 类的蛋白质条数；$\mathrm{FP_i}$ 表示非第 i 类亚细胞位置中的细胞凋亡蛋白质被错误的识别为第 i 类的蛋白质条数；FN_i 表示第 i 类亚细胞位置中细胞凋亡蛋白质被错误的识别为非 i 类的蛋白质条数。

5. 结果与讨论

5.1. 预测结果比较

本文分别从 mRNA 和氨基酸角度提取了不同的信息对细胞凋亡蛋白质的亚细胞位置进行预测。mRNA 方面分别利用了三阅读框下 3-mer mRNA 序列信息和 mRNA 二级结构－序列信息。根据三阅读框下 3-mer mRNA 序列信息可以构建一个 192 维的特征向量，mRNA 二级结构－序列信息可以构造成 32 维的特征向量，去掉部分特异性不强的信息，得到一个 23 维特征向量，将这两个特征融合后进行预测。基于 Jackknife 检验预测结果见表 3。由表 3 可以看出，采用 mRNA 单一信息时总体预测成功率分别达到 63.44%和 58.61%，而将 mRNA 序列与结构信息融合后总体预测成功率达到 65.56%，比三阅读框下 3-mer mRNA 序列信息提高了 2.12%，比 mRNA 二级结构－序列信息提高了 6.95%。结果表明，融合序列与结构信息后预测成功率有了提高，融合序列与结构信息能够更充分的反映出 mRNA 的特性。

氨基酸方面选用了物理化学性质，氨基酸黏性信息，进化信息三种特性。比较发现在物理化学性质中当 T = 50 时预测效果最佳，在氨基酸黏性特性中同样是当 $\kappa = 50$ 时预测效果最佳，进化信息中当 $\lambda = 1$ 时预测效果最佳。之后，将氨基酸的三种特性选取最优变量融合称为 AA hybrid，并对 AA hybrid 进行预测，预测结果见表 4。从表 4 可以看出，采用物理化学性质总体预测成功率达到 68.88%，氨基酸黏性信息总体预测成功率达到 70.69%，进化信息比前两者更高，达到 71%。将三种特性混合后总体预测成功率提高到 77.34%，比单一氨基酸特性最高提高了 8.46%。

由表 3，表 4 可以观察出，多特征融合后的总体预测成功率都要高于单特征的总体预测成功率，说明多特征融合可以更加全面的刻画细胞凋亡蛋白质。

将 mRNA 的特性与氨基酸的特性全部融合(hybrid)，利用融合后的特性进行预测。预测结果见表 5 所示。观察发现，同样是基于 Jackknife 检验，融合 mRNA 信息后预测效果更佳，比单独采用 mRNA 特性总体预测成功率提高了 16.62%，比只用氨基酸特性的总体预测成功率提高了 4.84%，这就表明 mRNA 对于细胞凋亡蛋白亚细胞定位预测具有一定作用。

Table 3. Prediction performance of mRNA feature parameter
表 3. mRNA 特性参数预测结果

Features (特征)	Localization (位置)	Sn (%)	Sp (%)	Acc (%)	MCC	OA (%)
3-mer mRNA sequence	Nuclear	65.98	83.76	78.55	0.49	63.44
	Membrane	72.53	87.92	83.69	0.60	
	Cytoplasmic	55.81	84.49	77.04	0.40	
	Mitochondrial	56.14	94.16	87.61	0.54	
structure-sequence mode	Nuclear	64.95	80.34	75.83	0.44	58.61
	Membrane	61.54	88.75	81.27	0.52	
	Cytoplasmic	53.49	82.45	74.92	0.36	
	Mitochondrial	50.88	92.34	85.20	0.46	
mRNA hybrid	Nuclear	64.95	85.47	79.46	0.50	65.56
	Membrane	73.63	89.58	85.20	0.63	
	Cytoplasmic	63.95	83.67	78.55	0.46	
	Mitochondrial	56.14	94.53	87.92	0.55	

Table 4. Prediction performance of AAs feature parameter
表 4. 氨基酸特性参数预测结果

Features (特征)	Localization (位置)	Sn (%)	Sp (%)	Acc (%)	MCC	OA (%)
physicochemical properties	Nuclear	70.10	80.34	77.34	0.48	68.88
	Membrane	85.71	96.67	93.66	0.84	
	Cytoplasmic	58.14	88.16	80.36	0.47	
	Mitochondrial	56.14	92.70	86.41	0.51	
Stickiness of AAs	Nuclear	76.29	82.05	80.36	0.56	70.69
	Membrane	76.92	93.33	88.82	0.72	
	Cytoplasmic	63.95	91.02	83.99	0.57	
	Mitochondrial	61.40	93.80	88.22	0.57	
PSSM	Nuclear	76.29	79.49	78.55	0.53	71.00
	Membrane	80.22	96.25	91.84	0.79	
	Cytoplasmic	62.79	86.53	80.36	0.49	
	Mitochondrial	59.65	97.81	91.24	0.67	
AA hybrid	Nuclear	79.38	88.03	85.50	0.66	77.34
	Membrane	90.11	95.00	93.66	0.84	
	Cytoplasmic	73.26	89.39	85.20	0.62	
	Mitochondrial	59.65	96.72	90.33	0.63	

Table 5. Hybrid feature under Jackknife test
表 5. 融合特征 Jackknife 检验

	Localization (位置)	Sn (%)	Sp (%)	Acc (%)	MCC	OA (%)
The jackknife test	Nuclear	86.60	90.60	89.43	0.75	82.18
	Membrane	92.31	96.67	95.47	0.89	
	Cytoplasmic	75.58	90.20	86.41	0.65	
	Mitochondrial	68.42	98.18	93.05	0.74	

Table 6. The predictive accuracies for the 262 dataset and the dent dataset
表 6. 262 数据集和独立测试集预测结果

	Localization (位置)	Sn (%)	Sp (%)	Acc (%)	MCC	OA (%)
The 262 dataset jackknife test	Nuclear	87.01	86.49	86.64	0.70	80.15
	Membrane	88.89	95.79	93.89	0.85	
	Cytoplasmic	72.06	92.27	87.02	0.66	
	Mitochondrial	66.67	98.16	92.75	0.73	
The Ident dataset	Nuclear	75.00	92.00	87,0	0.68	78.26
	Membrane	95.00	1.00	98.60	0.96	
	Cytoplasmic	94.00	80.00	84.10	0.67	
	Mitochondrial	33.00	98.00	87.00	0.46	

为了进一步评估算法对新序列的预测能力，本文进行了独立检验，分别从来自四个亚细胞位置的 331 条序列中随机挑选 20%，共 69 条序列构成独立测试集，命名为 Ident 数据集。剩余的 262 条序列构成 262 数据集。采用融合特征(hybrid)作为输入参数输入 SVM，262 数据集的 jackknife 检验和 Ident 数据集的独立检验预测结果列于表 6。观察表 6 发现，混合特征对 262 数据集具有较好的预测能力，总体预测成功率达到 80.15%。对于独立测试集预测成功率达 78.26%，反映出融合特征对未知细胞凋亡蛋白质也具有良好的预测能力。

5.2. 讨论

研究发现不同亚细胞位置细胞凋亡蛋白质的 mRNA 序列和结构具有一定的特异性。只采用氨基酸信息进行蛋白质亚细胞定位预测特征略显单一，将 mRNA 序列信息与二级结构信息融合后取得了更优的预测成功率。mRNA 局部二级结构与其功能密切相关，结构-序列模式中三联体参数从短片段的结构和碱基种类出发考虑 mRNA 执行功能时的局域性，揭示出 mRNA 与细胞凋亡蛋白质亚细胞位置的关系。采用支持向量机的方法在 Jackknife 检验下取得了良好的预测结果，进行独立测试后发现算法对新序列也有较好的预测能力，这也说明所选择的特征参数能够比较有效的区分不同亚细胞位置中的细胞凋亡蛋白质。本文综合考虑了氨基酸的物理化学特性，氨基酸的黏性，进化信息，mRNA 的序列信息和 mRNA 的二级结构信息，如果将更多的细胞凋亡蛋白质亚细胞位置的特征信息融合，更有效的提取序列中的蕴含的结构与功能信息，将对进一步提高细胞凋亡蛋白质的亚细胞定位预测有所帮助，也可能对进一步研究细胞凋亡蛋白质的功能提供一些理论依据。

致 谢

感谢国家自然科学基金(61361015)和教育部第 46 批留学回国人员科研启动基金的支持。

参考文献 (References)

[1] 屈二军, 胡建业, 陈兰英. 细胞凋亡与疾病研究进展[J]. 临床和实验医学杂志, 2008(8): 177-178.

[2] Zhirnov, O.P., Konakova, T.E., Wolff, T., *et al*. (2002) NS1 Protein of Influenza A Virus Down-Regulates Apoptosis. *Journal of Virology*, **76**, 1617-1625. http://dx.doi.org/10.1128/jvi.76.4.1617-1625.2002

[3] Reed, J.C. and Paternostro, G. (1999) Postmitochondrial Regulation of Apoptosis during Heart Failure. *Proceedings of the National Academy of Sciences of the USA*, **96**, 7614-7616. http://dx.doi.org/10.1073/pnas.96.14.7614

[4] Xue, C.H., Li, F., He, T., *et al*. (2005) Classification of Real and Pseudo microRNA Precursors Using Local Structure-Sequence Features and Support Vector Machine. *BMC Bioinformatics*, **6**, 310. http://dx.doi.org/10.1186/1471-2105-6-310

[5] Hofacker, I.L., Fontana, W., Stadler, P.F., *et al*. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte für Chemie/Chem Mon*, **125**, 167-188.

[6] Gao, Q.-B., Wang, Z.-Z., Yan, C. and Du, Y.-H. (2005) Prediction of Protein Subcellular Location Using a Combined Feature of Sequence. *FEBS Letters*, **579**, 3444-3448. http://dx.doi.org/10.1016/j.febslet.2005.05.021

[7] Lio, P. and Vannucci, M. (2000) Wavelet Change-Point Prediction of Transmembrane Proteins. *Bioinformatics*, **16**, 376-382. http://dx.doi.org/10.1093/bioinformatics/16.4.376

[8] Kawashima, S., Ogata, H. and Kanehisa, M. (2000) AAindex: Amino Acid Index Database. *Nucleic Acids Research*, **28**, 374. http://dx.doi.org/10.1093/nar/28.1.374

[9] Chou, K.-C. and Cai, Y.-D. (2006) Predicting of Protease Type in a Hybridization Space. *Biochemical and Biophysical Research Communications*, **339**, 1015-1020. http://dx.doi.org/10.1016/j.bbrc.2005.10.196

[10] Chou, K.-C. and Cai, Y.-D. (2006) Predicting Protein-Protein Interactions from Sequence in a Hybridization Space. *Journal of Proteome Research*, **5**, 316-322. http://dx.doi.org/10.1021/pr050331g

[11] Chou, K.-C. and Cai, Y.-D. (2004) Predicting Enzyme Family Class in a Hybridization Space. *Protein Science*, **13**, 2857-2863. http://dx.doi.org/10.1110/ps.04981104

[12] Amos-Binks, A., *et al*. (2011) Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery Using Re-Occurring Polypeptide Sequences. *BMC Bioinformatics*, **12**, 225. http://dx.doi.org/10.1186/1471-2105-12-225

[13] Gromiha, M.M. and Selvaraj, S. (2004) Inter-Residue Interactions in Protein Folding and Stability. *Progress in Biophysics & Molecular Biology*, **86**, 235-277. http://dx.doi.org/10.1016/j.pbiomolbio.2003.09.003

[14] Levy, E.D., De, S. and Teichmann, S.A. (2012) Cellular Crowding Imposes Global Constraints on the Chemistry and Evolution of Proteomes. *Proceedings of the National Academy of Sciences of the USA*, **109**, 20461-20466. http://dx.doi.org/10.1073/pnas.1209312109

[15] Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M. and Lundegaard, C. (2009) A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions. *BMC Structural Biology*, **9**, 51. http://dx.doi.org/10.1186/1472-6807-9-51

[16] Schaffer, A.A., Aravind, L., Madden, T.L., *et al*. (2001) Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-Based Statistics and Other Refinements. *Nucleic Acids Research*, **29**, 2994-3005. http://dx.doi.org/10.1093/nar/29.14.2994

[17] Chou, K.C. (2001) Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. *Proteins*, **43**, 246-255. http://dx.doi.org/10.1002/prot.1035

[18] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), **2**, 27.

[19] Chou, K.C. and Elrod, D.W. (1999) Protein Subcellular Location Prediction. *Protein Engineering*, **12**, 107-118. http://dx.doi.org/10.1093/protein/12.2.107

[20] Chou, K. and Zhang, C. (1995) Prediction of Protein Structural Classes. *Critical Reviews in Biochemistry and Molecular Biology*, **30**, 275-349. http://dx.doi.org/10.3109/10409239509083488