基于多序列特征提取的蛋白质相互作用预测

杜明宇[1,2]，张晓龙[1,2]

(1. 武汉科技大学 计算机科学与技术学院，湖北 武汉 430070;
2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室，湖北 武汉 430070)

摘 要：考虑到现有的基于序列的蛋白质相互作用预测方法均采用单一的特征提取方法，具有一定的局限性，提出一种方法。用元学习策略作为分类器融合策略，并集成多种蛋白质序列特征提取方法。在 10 702 对酿酒酵母蛋白质对数据集上，得到 97.28% 的预测精度，优于目前现有方法的平均水平，在独立测试集上同样具有优秀的表现，实验结果表明，该方法有效提高了蛋白质相互作用预测的准确率。

关键词：蛋白质-蛋白质相互作用；蛋白质序列；特征提取；支持向量机；分类器融合
中图法分类号：TP301 **文献标识号**：A **文章编号**：1000-7024 (2018) 01-0086-04
doi：10.16208/j.issn1000-7024.2018.01.016

Predicting protein-protein interactions from protein sequence based on multiple feature extractions

DU Ming-yu[1,2], ZHANG Xiao-long[1,2]

(1. College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430070, China;
2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan University of Science and Technology, Wuhan 430070, China)

Abstract: Considering that the existing methods for protein-protein interactions (PPIs) based on protein sequence use the single feature extraction and have certain limitations, a method based on multiple feature extraction for protein sequence was proposed. Experiments were carried out on the data set with 10 702 Saccharomyces cerevisiae protein pairs. Results show that the prediction accuracy of the proposed method reach 97.28%, which is superior to the average of the existing methods. On the indepen-dent test set, the proposed method also shows excellent performance, indicating that it effectively improves the accuracy of prediction of PPIs.

Key words: protein-protein interactions; protein sequence; feature extraction; support vector machine; classifier fusion

0 引 言

目前，蛋白质相互作用[1]的研究方法主要分为两大类：实验方法和计算方法。传统的实验方法费时费力，并且存在一些其它不可避免的缺陷，所以有些研究人员将目光转向计算方法，也已经提出过许多能有效的预测蛋白质-蛋白质相互作用计算方法：有基于基因组信息的[2]，也有利用蛋白质的结构信息和保守序列信息的方法[3]，还有通过研究已知的蛋白质的结构域的方法来预测蛋白质相互作用[4]。然而，这些方法需要的相应的蛋白质相关信息并不易取得，因此适用性并不普遍。基于序列的蛋白质相互作用预测方法，只涉及到蛋白质的序列信息，相较于其它方法需要的数据源更容易得到，发展迅速。Shen 等[5]提出三联体组合信息编码方法对蛋白质序列进行特征表示，达到了 83.9% 的预测准确率；Guo 等[6]进一步考虑到蛋白质序列内部存在更为复杂的相互作用，提出自协方差编码方式，使用基于高斯核函数的支持向量机 (SVM) 作为分类器，预测酿酒酵母蛋白质对的相互作用，得到 88.09% 的准确率；Zhou 等[7]提出了通过局部描述符来表示蛋白质序列，并结合支持向量机，预测蛋白质之间的相互作用。然而，上述方法

收稿日期：2016-11-07；修订日期：2016-12-20
基金项目：国家自然科学基金项目 (61273225、61502356)
作者简介：杜明宇 (1992-)，男，湖北仙桃人，硕士研究生，研究方向为生物信息处理、机器学习；张晓龙 (1963-)，男，湖北武汉人，博士，教授，研究方向为数据挖掘、机器学习。E-mail：1530230385@qq.com

在对蛋白质序列进行特征提取时，仅仅采用了一种特征提取方法，预测精确度有待提高。因此本文借鉴集成学习的思想[8]，通过融合多个分类器来集成多种特征提取方法，以期望提高从蛋白质序列中提取出的信息量来提高蛋白质相互作用预测的准确度。

1　特征提取

基于序列对蛋白质进行相互作用预测，首先得进行蛋白质序列的特征提取，也就是将长短不一的蛋白质序列编码成定长的特征向量，以便于接下来的模型构建。目前蛋白质序列特征提取方法有很多种，从中选取了具有代表性且编码方式有较大差异的 3 种序列表示方法，包括氨基酸组成编码、氨基酸理化属性编码和自协方差编码。

1.1　氨基酸组成编码

氨基酸组成（AAC）编码是最早被提出来的蛋白质序列的特征表达方法，它计算 20 种氨基酸中的每一类在蛋白质序列中出现的百分含量，于是一条蛋白质序列能被转化为一个 20 维的特征向量，并可以表示为下面的形式

$$X = (f_1, f_2, \cdots, f_{20}) \tag{1}$$

式中：$f_i = \frac{q_i}{\sum_{i=1}^{20} q_i}, i = 1, \cdots, 20$；其中，$f_i$ 为各种氨基酸的百分含量，q_i 为各种氨基酸的出现次数。

1.2　氨基酸理化属性组成编码

蛋白质序列是由若干氨基酸脱水缩合而成，各个氨基酸自身的理化属性对整体蛋白质存在着密切影响，于是在蛋白质序列信息的基础上，引入氨基酸的各种理化属性信息，氨基酸理化属性组成编码方法（PCC）被提出[9]。20 种氨基酸根据各种理化属性均被分为了 3 类[10]，具体分类情况见表 1。

表 1　20 种氨基酸依据 6 种理化属性的分类

理化属性	类别Ⅰ	类别Ⅱ	类别Ⅲ
疏水性	R,K,E,D,Q,N	G,A,S,T,P,H,Y	C,L,V,I,M,F,W
范德华体积	G,A,S,C,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
极性	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
极化率	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
电荷	K,R	A,N,C,Q,G,H,I,L,M,F,P,S,T,W,Y,V	D,E
二级结构	E,A,L,M,Q,K,R,H	V,I,Y,C,W,F,T	G,N,P,S,D

按照该分类一条蛋白质序列可以被分别转换为 6 条数值序列。将氨基酸以及与它相邻的两个氨基酸视为一个整体，每个氨基酸有 3 种类别，3 个氨基酸片段共 3* 3* 3 = 27 种排列。依次计算 6 条数值序列中这 3 个氨基酸片段以各种排列方式出现的次数，得到一个 6* 27 = 162 维的向量。该向量的取值与蛋白质序列长度有关，通常序列长度大的蛋白质序列编码出的特征向量值较大，按 Z-score 标准化规则对其进行数据标准化处理，作为该蛋白质序列对应的特征向量。

1.3　自协方差编码

Guo 等完整地提出了使用自协方差（AC）来描述蛋白质序列的方法，其中首要涉及到氨基酸的各种理化属性，选取其中得到比较广泛认可的 6 种理化属性，包括疏水性、极性、极化率、侧链体积、溶剂可及表面积和侧链的净电荷指数[11]，对其原始测定值按式(2) 进行标准化处理

$$P'_{i,j} = \frac{P_{i,j} - P_j}{S_j} \tag{2}$$

式中：$P_{i,j}$为第 i 种氨基酸的第 j 种理化属性的测定值，P_j 和 S_j 为 20 种氨基酸的第 j 种理化属性的平均值和标准差。

给定一条蛋白质序列，对特定的属性，每个氨基酸被替换为相对应的标准化后的属性数值，然后通过式(3) 所示的自协方差描述符编码为 $lag* j$ 维度的特征向量 $AC_{lag,j}$

$$AC_{lag,j} = \frac{1}{n - lag}\sum_{i=1}^{n-lag}\left(X_{i,j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right) \times \left(X_{(i+lag),j} - \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right) \tag{3}$$

式中：$X_{i,j}$是长度为 n 的蛋白质序列中第 i 个氨基酸对应的第 j 种属性数值，lag 为滑动窗的宽度。在本实验中，参数 lag 和 j 分别取 30 和 6，因此一条蛋白质序列最终被编码成 30* 6 = 180 维向量。

上述编码方法均是对单条蛋白质序列而言，即仅将单条蛋白质序列编码为相应的特征向量，而蛋白质相互作用的预测是基于蛋白质对，因此需要将两条蛋白质序列分别编码后对应的特征向量结合起来，作为代表这对蛋白质对的特征向量。这里采用最普遍的矢量拼接方式，即 $a \oplus b$。以氨基酸组成编码为例，蛋白质 A 和 B 分别被编码为 20 维向量（$m1$，$m2$，$m3$，…，$m20$）和（$n1$，$n2$，$n3$，…，$n20$），则蛋白质对 A-B 可以表示为向量（$m1$，$m2$，$m3$，…，$m20$，$n1$，$n2$，$n3$，…，$n20$）。

2　实验设计

2.1　数据集

本文采用的 PPIs 数据来自 DIP 数据库[12] 中酿酒酵母核心数据集 DIP_ 20151029 版本。其中含有 5399 对相互作用蛋白质对。去掉包含序列长度低于 50 的蛋白质对，得到 5351 对相互作用的蛋白质对，作为本实验的正集。

由于当前收录无相互作用的蛋白质的数据库还比较少，因此需要人工构造非相互作用蛋白质对。处于细胞中不同亚细胞位置的蛋白质之间不会发生相互作用，将亚细胞位置信息不同的蛋白质随机配对，得到的蛋白质对认为其不相互作用。蛋白质的亚细胞信息可以从 UniProt 数据库[13]获得，从 UniProt 数据库拿到的酿酒酵母的数据有 6721 条，去掉其中不含亚细胞位置信息和对应 DIP 数据库蛋白质编号的数据，将余下的数据按照亚细胞位置（Cytoplasm、Nucleus、Mitochondrion、Endoplasmic reticulum、Golgi apparatus、Peroxisome、Vacuole）分成 7 组，将位于不同分组的蛋白质随机配对，并去掉与正集中重复的蛋白质对，得到若干不存在相互作用的蛋白质对，为了保证数据的平衡性，取其中的 5351 条数据，作为本实验的最终负样本。

2.2 模型构造

用于训练和测试的最终实验样本集包含 10 702 对蛋白质对，其中存在相互作用的占一半，没有相互作用的占一半。由于较大的正负样本容量和建模计算量，采用简单交叉验证来验证改模型的性能。从正负样本中随机抽取 7000 条数据作为训练集用于模型的构建，剩余的样本作为测试集进行模型评估。训练集中的样本是从实验样本集从随机抽样得到，单次使用简单交叉验证得到的估计结果是不够稳定可靠的，所以将实验过程重复 10 次，讲得到的 10 组模型的各个指标的平均值作为评估结果。此外，蛋白质相互作用预测是一个二分类问题，即预测结果是存在相互作用或者不可能发生相互作用，分类器的选择上选用支持向量机。因为蛋白质序列编码出的特征向量与两个蛋白质是否相互作用不是一个简单的线性关系，支持向量机的核函数采用高斯核函数，参数 C 使用默认的值 1（C 过大或者过小，分类器的泛化能力变差，默认值满足我们的条件）。模型的构造基于 R 语言平台实现，具体结构如图 1 所示。

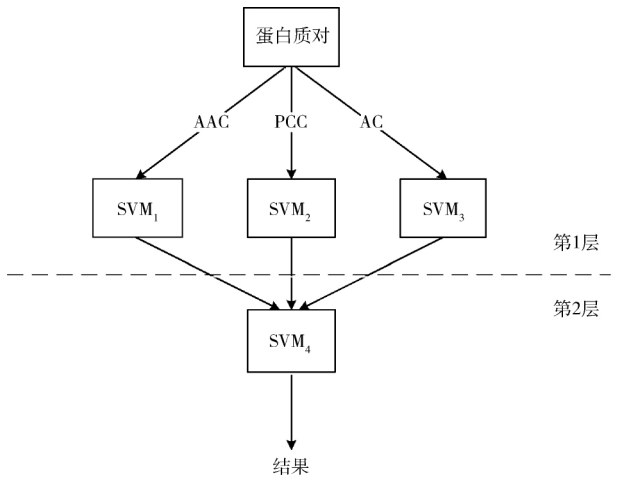

图 1 基于多序列特征提取的蛋白质相互作用预测模型

蛋白质序列对分别按照氨基酸组成（AAC）、氨基酸理化属性组成（PCC）、自协方差编码（AC）3 种特征提取方法处理，依次被编码成 40、324、360 维度的特征向量，然后在 3 个互相独立的支持向量机 SVM_1、SVM_2、SVM_3 中经过模型判别，得到各自预测的结果 C_1、C_2、C_3。将来自 3 个分类器的预测结果拼接成向量（C_1，C_2，C_3），作为第二层的支持向量机 SVM_4 的输入，得到的结果即为我们的整个模型的预测结果。

2.3 性能评估

模型的预评估采用敏感度（SN）、特异性（SP）和准确率（AC）这 3 个指标来评价，它们的计算公式如下所示

$$SN = \frac{TP}{TP + FN} \tag{4}$$

$$SP = \frac{TN}{TN + FP} \tag{5}$$

$$AC = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

其中：TP（true positive）表示真实相互作用对的数量，FP（false positive）表示假的相互作用数量，FN（false negative）表示假的非相互作用对的数量，TN（true negative）表示真实的非相互作用的数量。

3 结果及分析

3.1 预测结果

本文提出的基于多特征提取方法在测试集上的 10 次实验结果见表 2。从表 2 可以看出，10 次实验的准确率均大于 97%，且在 1% 之内波动，平均敏感度、平均特异性和平均准确率分别达到 97.24%、98.09%、97.66%，说明模型在 10 702 对酿酒酵母蛋白质数据集能取得不错的预测效果，且具有良好的稳定性。

表 2 基于多特征提取方法的预测结果

实验次数	*TN*	*FN*	*FP*	*TP*	*SN*/%	*SP*/%	*AC*/%
1	1766	57	30	1849	97.01	98.33	97.65
2	1803	48	34	1817	97.43	98.15	97.78
3	1825	51	41	1785	97.22	97.80	97.51
4	1860	43	40	1759	97.61	97.89	97.76
5	1813	44	40	1805	97.62	97.84	97.73
6	1833	52	38	1779	97.16	97.97	97.57
7	1810	53	35	1804	97.15	98.10	97.62
8	1788	54	27	1833	97.14	98.51	97.81
9	1786	57	34	1825	96.97	98.13	97.54
10	1825	54	34	1789	97.07	98.17	97.62
平均值	1810.9	51.3	35.3	1804.5	97.24	98.09	97.66

3.2 与其它方法的对比

为了验证基于多特征提取方法的优劣，分别使用氨基酸组成、氨基酸理化性质组成和自协方差 3 种单种特征提取方法对蛋白质序列进行编码，基于支持向量机训练得到相应的分类器，3 种单种特征提取方法和基于多特征提取方法得到的分类器的平均性能表现见表 3。

表 3 本文提出的多特征提取方法和其它方法在实验样本上的平均表现

方法	*SN*/%	*SP*/%	*AC*/%
ACC	90.86	90.92	90.89
PCC	95.46	94.46	94.96
AC	96.56	93.81	95.19
本文方法	97.24	98.09	97.66

可以看到，使用氨基酸组成作为特征提取方法，灵敏度、特异性和准确率分别为 90.86%、90.92%、90.89%。而使用氨基酸理化属性组成来编码序列时，3 项指标均提高了 4% 到 5% 左右。这是因为氨基酸组成仅仅考虑到蛋白质序列中的氨基酸的组成信息，而氨基酸属性组成则涉及到氨基酸的各种属性信息，说明它能比氨基酸组成编码更好地反映相互作用蛋白质对之间的相互作用。自协方差编码方式，将蛋白质内部存在的相互作用考虑在内，并且包含了氨基酸残基的位置信息，与氨基酸理化属性作为特征提取方法相比，两者训练得到的分类器在实验样本上的表现相差 1% 以内。

本文提出的多特征提取方法模型融合了前 3 种特征提取方法，在本实验的数据集上，灵敏度上高于前三者中表现最好的自协方差方法，达到了 97.42%，特异性和准确率上高于前三者中表现最好的氨基酸组成方法，分别为 98.09% 和 97.66%，说明集成多种序列特征提取方法从蛋白质序列中提取出了充足且不冗余的信息，各方面表现均高于单种特征提取方法，有效的提高了蛋白质相互作用预测的精度。

3.3 独立测试集表现

为了进一步的研究本方法在蛋白质相互作用预测上的有效性，从 DIP 数据库中选取了其它 6 种不同物种的蛋白质数据作为独立的测试集，对模型的泛化性能进行验证，具体结果见表 4。从表 4 可以看出，6 个物种的数据集中有 4 个物种的预测精度超过了 90%，其它两个物种的预测精度也接近 90%，平均预测精度达到了 93.35%，表明基于多序列特征提取预测蛋白质相互作用有在未知的数据上也能表现的很好，具有较好的外推能力，进一步地说明了该方法是有效的对模型的泛化性能进行了的评估。

表 4 独立测试集表现

独立测试集	存在相互作用的蛋白质对数	预测正确率/%
C. elegans	3946	94.07
E. coli	12 212	89.04
H. pylori	1420	95.98
H. sapiens(Human)	6879	88.24
M. musculus(house mouse)	2352	96.59
R. norvegicus(Norway rat)	546	96.15
Total average	27 355	93.35

4 结束语

本文提出了一种集成多种蛋白质序列特征提取方法的模型来预测蛋白质对的相互作用，该方法利用集成学习的思想，使用多个支持向量机融合了氨基酸组成、氨基酸理化信息、自协方差 3 种不同的编码方法，实验结果表明，我们的方法是正确且可行的。但是也具有一定的局限性，因为涉及到自协方差编码，进行预测的蛋白质序列长度不能低于 30。接下来，在分类器的选择上，可以尝试使用其它机器学习算法，或是使用更好的特征提取方法来进一步完善本文提出模型。

参考文献:

[1] ZHANG Changsheng, LAI Luhua. Protein-protein interaction: Prediction, design, and modulation [J]. Acta Physico-Chimica Sinica, 2012, 28 (10): 2363-2380 (in Chinese). [张长胜，来鲁华. 蛋白质相互作用预测、设计与调控 [J]. 物理化学学报，2012，28 (10): 2363-2380.]

[2] Emamjomeh A, Goliaei B, Torkamani A, et al. Protein-protein interaction prediction by combined analysis of genomic and conservation information [J]. Genes & Genetic Systems, 2014, 89 (6): 259-272.

[3] Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale [J]. Nature, 2012, 490 (7421): 556-560.

[4] Binny PS, Saha S, Anishetty R, et al. A matrix based algorithm for protein-protein interaction prediction using domain-domain associations [J]. Journal of Theoretical Biology, 2013, 326 (23): 36-42.

[5] Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information [J]. Procee-dings of the National Academy of Sciences of the United States of America, 2007, 104 (11): 4337-4341.

(下转第 254 页)

[2] Lai Chung-Liang, Huang Ya-Ling, Liao Tzu-Kuan, et al. A Microsoft Kinect-based virtual rehabilitation system to train balance ability for stroke patients [C] //International Conference on Cyberworlds. IEEE, 2015: 54-60.

[3] Pei Wei, Xun Guanghua, Li Min, et al. A motion rehabilitation self-training and evaluation system using Kinect [C] //13th International Conference on Ubiquitous Robots and Ambient Intelligence. IEEE, 2016: 353-357.

[4] Borghese NA, Pirovano M, Lanzi PL, et al. Computational intelligence and game design for effective at-home stroke rehabilitation [J]. Games for Health Journal, 2013, 2 (2): 81-88.

[5] QU Chang, DING Chen, WANG Junze, et al. A method to measure the range of motion of human upper limbs based on Kinect somatosensory interaction technology [J]. Chinese Journal of Biomedical Engineering, 2014, 33 (1): 16-20 (in Chinese). [瞿畅, 丁晨, 王君泽, 等. 基于 Kinect 体感交互技术的上肢关节活动度测量方法 [J]. 中国生物医学工程学报, 2014, 33 (1): 16-20.]

[6] XIN Yizhong, XING Zhifei. Human action recognition method based on Kinect [J]. Computer Engineering and Design, 2016, 37 (4): 1056-1061 (in Chinese). [辛义忠, 邢志飞. 基于 Kinect 的人体动作识别方法 [J]. 计算机工程与设计, 2016, 37 (4): 1056-1061.]

[7] HUANG Jian. Research and design of the rehabilitation platform based on the Kinect motion of acquisition [D]. Chengdu: University of Electronic Science and Technology of China, 2013 (in Chinese). [黄健. 基于 Kinect 系统运动采集的康复训练平台的研究与设计 [D]. 成都: 电子科技大学, 2013.]

[8] Rodrigo Ibanez, Álvaro Soria, Alfredo Teyseyre, et al. Easy gesture recognition for Kinect [J]. Advances in Engineering Software, 2014, 76: 171-180.

[9] Li Nianfeng, Dai Yinfei, Wang Rongquan, et al. Study on action recognition based on Kinect and its application in rehabilitation training [C] //IEEE 5th International Conference on Big Data and Cloud Computing. IEEE, 2015: 265-269.

[10] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, et al. Real-time human pose recognition in parts from single depth image [C] // IEEE Conference on Computer Vision and Pattern Re-cognition. IEEE, 2011: 1297-1304.

[11] SHANG Huaqiang. Simulation study of virtual movement based on Kinect [D]. Hangzhou: Hangzhou Electronic Scie-nce and Technology University, 2012 (in Chinese). [尚华强. 基于 Kinect 的虚拟人物动作仿真研究 [D]. 杭州: 杭州电子科技大学, 2012.]

[12] Jarrett Webb, James Ashley. Beginning Kinect programming with the Microsoft Kinect SDK [M]. New York: APress, 2012: 93-94.

(上接第 89 页)

[6] 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 Research, 2008, 36 (9): 3025-3030.

[7] Zhou YZ, Gao Y, Zheng YY. Prediction of protein-protein interactions using local description of amino acid sequence [M] //Communications in Computer & Information Science, 2011: 254-262.

[8] Mallipeddi R, Suganthan PN, Pan QK, et al. Differential evolution algorithm with ensemble of parameters and mutation strategies [J]. Applied Soft Computing, 2011, 11 (2): 1679-1696.

[9] LI Juanjuan. Protein-protein interaction prediction based on computational intelligence [D]. Jinan: University of Jinan, 2014 (in Chinese). [李娟娟. 基于多特征融合和集成的蛋白质相互作用预测 [D]. 济南: 济南大学, 2014.]

[10] Li ZR, Lin HH, Han LY, et al. Profeat: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence [J]. Nucleic Acids Research, 2015, 34 (Web Server issue): 32-37.

[11] BI Jingye. Researches on the protein-protein interaction prediction method based on the sequence [D]. Taiyuan: Shanxi University, 2013 (in Chinese). [毕敬业. 基于序列的蛋白质相互作用预测方法研究 [D]. 太原: 山西大学, 2013.]

[12] Salwinski L, Miller CS, Smith AJ, et al. The database of interacting proteins: 2004 update [J]. Nucleic Acids Research, 2004, 32 (sup1): 449-451.

[13] Consortium UP. Update on activities at the universal protein resource (UniProt) in 2013 [J]. Nucleic Acids Research, 2013, 41 (Database issue): D43.